ESTUDIO DE LA DESERCIÓN ESTUDIANTIL EN LAS UNIDADES ACADÉMICAS EN PROGRAMAS DE PREGRADO DE LA MODALIDAD PRESENCIAL DE LA UNIVERSIDAD DEL TOLIMA MEDIANTE ANÁLISIS ENVOLVENTE DE DATOS - DEA Y REGRESIÓN BETA

NUBIA BERMÚDEZ VARÓN

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍA INDUSTRIAL
MAESTRÍA EN INVESTIGACIÓN OPERATIVA Y ESTADÍSTICA
PEREIRA - RISARALDA
Agosto de 2017

ESTUDIO DE LA DESERCIÓN ESTUDIANTIL EN LAS UNIDADES ACADÉMICAS EN PROGRAMAS DE PREGRADO DE LA MODALIDAD PRESENCIAL DE LA UNIVERSIDAD DEL TOLIMA MEDIANTE ANÁLISIS ENVOLVENTE DE DATOS - DEA Y REGRESIÓN BETA

Tesis presentada como requisito para optar al título de Magíster en Investigación Operativa y Estadística

NUBIA BERMÚDEZ VARÓN

DIRECTOR:
Mg. ALFONSO SÁNCHEZ HERNÁNDEZ

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍA INDUSTRIAL
MAESTRÍA EN INVESTIGACIÓN OPERATIVA Y ESTADÍSTICA
PEREIRA - RISARALDA

Agosto de 2017

DEDICATORIA

A mi padre Gentil y a mi madre Blanca Oliva por todo su apoyo y amor incondicional. A mis hijas Sandra Milena, Yesica Daniela y María Camila porque son el motor de mi vida y a mi esposo Ronald, mi compañero con todo mi amor.

AGRADECIMIENTOS

A mi Padre Celestial por su presencia y amor incondicional en nuestras vidas.

A mis padres por su dedicación y amor.

A mis hijas por su apoyo permanente.

A mi director de tesis, Alfonso Sánchez Hernández, profesor de planta de la Universidad del Tolima, por su apoyo permanente en este proceso de formación.

A la Universidad Tecnológica de Pereira y la Facultad de Ingeniería Industrial, quienes aunaron esfuerzos para permitirme terminar este postgrado que hace parte de mi proyecto de vida.

Tabla de Contenido

Índice de figuras

Índice de tablas

RESUMEN

La deserción estudiantil a nivel nacional, se ha convertido en una problemática producto de políticas de estado, siendo la principal causa, la ampliación de cobertura. A raíz de esta problemática el Ministerio de Educación Nacional - MEN viene buscando estrategias que permitan entender y evitar que los estudiantes universitarios deserten de las instituciones.

En el presente trabajo, se realizó un estudio de deserción estudiantil con variables socioeconómicas, académicas, individuales y financieras de nueve unidades académicas de la modalidad presencial de la universidad del Tolima, con un enfoque metodológico cuantitativo y cualitativo. Para ello, se realizó análisis envolvente de datos y análisis de regresión beta.

Palabras claves: Análisis envolvente de datos, regresión beta y deserción estudiantil.

1. INTRODUCCIÓN

La deserción estudiantil, es un fenómeno difícil de analizar, debido a que en ella intervienen demasiados factores que hacen que ésta se revise desde diferentes temáticas para su estudio. En este trabajo, se pretende realizar un aporte serio, responsable y consecuente con la realidad. Ante esta problemática no podemos ser actores pasivos de la misma, por el contrario, se deben tener en cuenta herramientas necesarias para que la deserción no continúe de manera descontrolada y afecte a los mismos estudiantes, a sus familias, al Estado y a la comunidad en general; cuyos recursos económicos, sociales y culturales involucrados son muy importantes. No se pueden permitir acciones que vayan en dirección contraria, es decir, mantenerla. Por el contrario la misión es evitarla y controlarla con indicadores mínimos. Aunque la deserción, es un asunto individual, único y propio de la decisión personal, es importante conocer cuáles son sus causas, sus circunstancias internas o externas y cuáles serían sus posibles soluciones o al menos plantearlas, para que la comunidad educativa en general (caso puntual Universidad del Tolima), se entere de esta situación y por ende genere estrategias que aseguren una mayor permanencia de sus estudiantes. El estudio de deserción, debe responder a dichas estrategias que permitan su disminución en los programas académcos de la modalidad presencial.

Adicionalmente, se realizó la revisión de antecedentes que incluyen universidades públicas y privadas del país, y del exterior. Las mismas han formulado políticas de gobierno para evitar este flagelo. En el caso de Colombia, actualmente se definió una politica denominada *Permanencia y graduación*, que busca por parte del Ministerio de Educación Nacional, el acompañamiento a las instituciones de educación superior para implementar las ocho componentes, a saber: Posicionamiento y Formalización, Cultura de la Información, Mejoramiento de Calidad Académica, Trabajo conjunto con Instituciones de Educación Media, Gestión de recursos, Fortalecimiento programas de fomento a la permanencia, Compromiso núcleo familiar y Trabajo colaborativo y en red. El Ministerio de Educación Nacional, conjuntamente con la Universidad de los Andes, han propuesto modelos de duración o de análisis de supervivencia que consiste en hacer seguimiento al estudiante desde el momento de su ingreso hasta que ocurre el evento, que en este caso es la deserción, pero no profundiza en los determinantes principales como son los factores financieros, socioeconómicos, institucionales, orientación vocacional y profesional.

Así mismo, otros estudios, como el caso de la Universidad Nacional, han trabajado Modelos Lineales generalizados y a nivel latinoamericano se han trabajado en Modelos de Regresión Logit, Probit, Análisis Discrimínate, entre otros.

Sin embargo, no han abordado de manera concreta nuevas estrategias como la asignación de recursos físicos, financieros, el análisis en las áreas de conocimiento en las cuales están inmersos los estudiantes que pueden ser apoyados por las instituciones y así, evitar su retiro en los primeros semestres de los programas académicos de las uni-

versidades del país, como se propone en el desarrollo de este trabajo.

Dado lo anterior, este proyecto pretende proponer un estudio basado en la aplicación de la técnica de Análisis Envolvente de Datos, que incluya algunos de los factores determinantes de la deserción expuestos anteriormente, al igual que integrar en él, los recursos de talento humano, de infraestructura y financieros, combinándolo con Análisis de Regresión Beta con las variables que están contempladas en el análisis DEA; que permitan entender el problema de la deserción estudiantil y como guía para el diagnóstico, diseño de acciones y evaluación de las mismas, cuyo propósito principal es el desarrollo de este trabajo.

2. PLANTEAMIENTO DEL PROBLEMA

La Universidad del Tolima, tiene como misión fomentar "el desarrollo de capacidades humanas para la formación integral permanente, apoyada en valores éticos de tolerancia, respeto y convivencia mediante la búsqueda incesante del saber, la producción y la apropiación y divulgación del conocimiento en los diversos campos de la ciencia, el arte y la cultura, desde una perspectiva inter y transdisciplinar, como aporte al bienestar de la sociedad, al ambiente y al desarrollo sustentable de la región, la Nación y el mundo"[1].

Que además, es el ente de educación superior más importante en la región que debe velar porque la educación llegue a todos los niveles de la sociedad. Una de las preocupaciones más importantes, ha sido la deserción estudiantil, que a nivel nacional está en un 46.1 % y que a pesar de estar por debajo del promedio nacional 46 %, sigue siendo un factor desestabilizante para el progreso de la región, como se evidencia en la pagina web del MEN.

La Universidad del Tolima y en general las instituciones de educación superior, no cuentan con una herramienta eficiente que permita optimizar los recursos disponibles en pro de la permanencia estudiantil.

Con la información que posee la Universidad del Tolima, se puede identificar una correlación entre las eficiencias calculadas con DEA y los niveles de deserción estudiantil de pregrado presencial en las unidades académicas, apoyados en Regresión Beta.

3. OBJETIVOS

3.1. Objetivo General

Desarrollar un estudio de deserción estudiantil, identificando su correlación con las Eficiencias calculadas mediante el Análisis Envolvente de Datos –DEA y Regresión Beta, en las nueve unidades académicas, que poseen programas de pregrado en la modalidad presencial en la Universidad del Tolima –UT.

3.2. Objetivos Específicos

Elegir un modelo DEA adecuado para aplicarlo a la deserción, en las unidades académicas de la UT.

Determinar las eficiencias relativas de las DMUs (unidades académicas) del modelo de deserción estudiantil para la optimización de recursos.

Encontrar un modelo de regresión Beta que explique la deserción en términos de las eficiencias y otras variables que no quedaron incluidas en el modelo DEA.

Realizar un análisis de diagnóstico que permita ls validación y consideración un modelo de regresión Beta, que permita identificar situaciones reales en el caso de la deserción estudiantil.

4. JUSTIFICACIÓN

El Ministerio de Educación Nacional define la educación como: "uno de los instrumentos más importantes con los que puede contar un país para asegurar su desarrollo humano y social"(MEN, 2010) [5]; inicialmente el objetivo de la política pública estaba encaminada a la ampliación de la cobertura, la cual generó el aumento de la deserción estudiantil en educación superior en Colombia lo que se convirtió en política pública en Colombia, fijándose metas como es disminuir la deserción por cohorte al 40 % en el año 2010 y al 25 % en el año 2019. Es así, como el MEN inicia un trabajo en conjunto con la Universidad de los Andes en el año 2004, donde se seleccionaron unas universidades por su tamaño y complejidad como muestra piloto para la implementación de la herramienta del Sistema de Prevención y Análisis a la Deserción en las Instituciones de Educación Superior –SPADIES, en la cual estaba inmersa la Universidad del Tolima.

Posteriormente, en el año 2010 se realizó una convocatoria para asignar recursos a las universidades del país apoyando el tema de permanencia estudiantil, como estrategia para evitar la deserción. La Universidad del Tolima inicio el proceso de seguimiento a los estudiantes de los niveles 1 y 2, cuyo resultado fue el diagnostico en su primera fase a estos estudiantes, evidenciado las dificultades en Lectoescritura, Matemáticas, Física y Química; que permitió que se iniciara el acompañamiento con monitores, los cuales cursaban entre 8o. y 10o. semestre, para la realización de cursos nivelatorios en estas asignaturas.

Adicionalmente, el MEN en el año 2010 invita al británico Ormond Simpson experto en el tema de deserción estudiantil al Foro Internacional sobre Permanencia Estudiantil en Educación Superior, para que compartiera las experiencias que han sido implementadas en la Universidad Abierta del Reino Unido (OU), basándose en focalizar su estrategia en aumentar los estímulos a los estudiantes de primer semestre para que terminen la carrera mediante cuatro actividades puntuales: seleccionar adecuadamente el programa académico, identificar los estudiantes con deficiencias, ofrecer un contacto proactivo y apoyo exógeno. La OU utiliza un modelo estadístico probabilístico de deserción, basado en características individuales como sexo, edad, nivel de dificultad del curso, la cantidad de cursos tomados y ocupación laboral. De igual manera, los estudios realizados por el MEN, han identificado que el determinante principal de la deserción es la dimensión académica y continúan en orden descendente los factores financieros, socioeconómicos, institucionales, orientación vocacional y profesional. Los modelos utilizados han sido la Regresión Logit, Probit, Análisis Discrimínante, Modelos de duración o de Análisis de Sobrevivencia, este último es el que ha tomado más fuerza para realizar estudios de deserción estudiantil, debido a que permite realizar análisis dinámicos del fenómeno.

Es así, como este proyecto busca como resultado eficiencias entre las diferentes unidades académicas aplicando el Análisis Envolvente de Datos –DEA con Análisis de Regresión Beta.

5. DISEÑO METODOLÓGICO

El presente proyecto, se realizará con una búsqueda bibliográfica exhaustiva asociada a la deserción estudiantil de las universidades públicas del país, con cada uno de los pasos para alcanzar los objetivos específicos:

Objetivo 1. Elegir un modelo DEA adecuado para aplicarlo a la deserción, en las unidades académicas de la UT.

Paso 1: Recolectar los datos de las unidades académicas que servirá de insumo para las variables de entrada y salida de un modelo DEA. Se tiene identificadas las siguientes variables: número de docentes de planta, número de docentes catedráticos, recursos asignados a cada unidad, número de estudiantes, programas académicos de cada unidad, número de funcionarios administrativos de cada facultad, espacios físicos asignados según horarios de clase. Productos de investigación de cada facultad.

Paso 2: Seleccionar cuáles variables serán las de entrada y cuáles serán las de salida para el modelo DEA adecuado para el respectivo análisis.

Paso 3: Revisar los diferentes modelos DEA en la literatura y elegir el más adecuado para el estudio de deserción en la presente tesis.

Objetivo 2. Determinar las eficiencias relativas de las DMUs (unidades académicas) del modelo de deserción estudiantil para la optimización de recursos.

Paso 1: Utilizar el modelo DEA seleccionado para encontrar las eficiencias relativas de cada una de las DMUs determinadas para este propósito.

Paso 2: Interpretar los resultados anteriores y posibles recomendaciones a las diferentes facultades.

Objetivo 3. Encontrar un modelo de regresión Beta que explique la deserción en términos de las eficiencias y otras variables que no quedaron incluidas en el modelo DEA.

Paso 1: Identificar las variables independientes tipo explicativas y no incluidas en el estudio DEA debido la naturaleza de las mismas (cualitativas por ejemplo).

Paso 2: Plantear un modelo de regresión para las variables regresoras que explican la deserción como variable respuesta. En este caso como variable de respuesta se manejan dos estados, alta deserción, baja deserción.

Paso 3: Encontrar los parámetros de la regresión propuesta y su validación.

Paso 4: Revisar otros tipos de regresiones convenientes para el estudio.

Paso 5: Examinar la validez de los resultados, mediante el análisis de Influencia del modelo estimado.

Paso 6: Comparar los diferentes modelos, con el fin de verificar cual explica mejor el fenómeno de deserción. Realizar un análisis de diagnóstico para medir la calidad de ajuste del modelo y análisis de sensibilidad del mismo.

Objetivo 4: Realizar un análisis de diagnóstico que permita ls validación y consideración un modelo de regresión Beta, que permita identificar situaciones reales en el caso de la deserción estudiantil.

Paso 1: Determinar un modelo que ajuste los datos.

Paso 2: Modelar las variables independientes no paramétrica.

Paso 3: Determinar la familia distribucional a la que pertenece la variable respuesta.

6. MARCO TEÓRICO

6.1. Modelos DEA convencionales

El Análisis Envolvente de datos, Data Evelopment Analysis (DEA) es una técnica no paramétrica usada para evaluar la eficiencia relativas de un conjunto de DMU (Decision Making Units) DEA, con fundamentos de programación lineal que busca eficiencia tanto en los insumos como en los productos, ya sea en empresas manufactureras, bienes o servicios, que indaga la solución eficiente en un problema de investigación de operaciones. El objetivo fundamental de DEA, es optimizar la eficiencia relativa de cada DMU, para establecer una frontera de eficiencia, usando el criterio de eficiencia de Pareto. DEA considera que la j-ésima DMU es eficiente, si elabora más unidades de alguno de los productos fabricados, manteniendo la producción de los otros, usando las mismas entradas o si puede generar las mismas salidas, utilizando una menor cantidad de, al menos, una entrada. La frontera eficiente está conformada por aquellos DMU eficientes. Después de obtenida esta frontera, se evalúa la eficiencia de cada DMU que no pertenezca a ésta, asumiendo que no existen perturbaciones aleatorias. La idea es comparar cada DMU no eficiente con aquella que lo sea y, además, tenga una técnica de producción similar. En general, la unidad con la que se comparan la DMUS ineficientes es una combinación lineal de las DMUS eficientes. Estas unidades ficticias reciben el nombre de grupo de referencia.

Ésta temática surge con la dirección del profesor Cooper en la tesis de Edwardo Rhodes, en la Carnegie Mellon University, con el trabajo enfocado a la evaluación de programas para estudiantes con dificultades en el aprendizaje en escuelas en E.E.U.U. apoyados por el Gobierno Federal, programa denominado “Follow Through”.

Tal como lo señala Charnes, Cooper y Rhodes, fue Farrell con su artículo “The Measurement of Productive Efficiency”(publicado en 1957 por la revista Journal of the Royal Statistical Society) [3], el autor más influyente en temas relacionados con la medición de la productividad y la eficiencia.

Farrell (1951) [6] propuso una medida de la eficiencia de una empresa dividida en dos componentes: eficiencia técnica y eficiencia de asignación. Éstas se combinan en una medida única de eficiencia global: la denominada eficiencia económica. No obstante, para ser capaces de calcular una medida de eficiencia, es necesario conocer previamente la forma explícita de la función de producción. Dado que, en la práctica, la frontera de producción nunca es conocida, Farrell [6] sugirió que esta función podría ser estimada a partir de una muestra de datos usando, alternativamente, una tecnología no paramétrica lineal a trozos, o bien, una función de producción paramétrica. Estas ideas condujeron, décadas más tarde, a dos metodologías claramente diferenciadas: el DEA y las fronteras estocásticas, respectivamente. Mientras que el DEA utiliza herramientas de la programación matemática, la aproximación a la medición de la eficiencia a través de fronteras

estocásticas recurre a técnicas de carácter puramente estadístico–econométrico [1].

Inicialmente para comprender los modelos básicos de la herramienta no paramétrica–DEA, debe tenerse claros los conceptos de eficiencia enfocados a la producción e insumos, así:

Productividad: El primer trabajo sobre pro- ductividad se remonta en los trabajos de Farrell (1957) [6]. Según Soto y Arenas (2010) [2], La productividad debe ser entendida como la relación entre el nivel de producción final obtenido y los recursos o insumos necesarios para lograrlo. En 1950 la organización para la Cooperación Económica Europea se refiere a la productividad de factores como "El cociente que se obtiene al dividir la producción entre uno de los factores de producción ".

Eficiencia: resultados obtenidos (outputs) y los recursos utilizados (inputs), en ese orden de ideas la eficiencia, será en cualquier caso una magnitud multidimensional.

Diferencias entre eficiencia y productividad: El cálculo de la productividad para una empresa es poco ilustrativo (un valor aislado de productividad, no es autónomo para explicar si es bueno o malo), a no ser que se haga referencia a otras empresas, respecto al aprovechamiento que se hace de los recursos (inputs) empleados en la producción de los productos (outputs); por lo que es necesario expresarla como eficiencia.

La maximización del beneficio exige que la organización tome correctamente las tres siguientes decisiones:

De entre todos los niveles de producción posibles, debe elegir el output (producto) que maximice el beneficio.

De entre todas las combinaciones de inputs (insumos), que sirven para producir el nivel de outputs (productos) anterior, debe elegirse aquella combinación de inputs que minimiza el costo de producción.

La organización debe producir el output elegido con la cantidad mínima de inputs posible, es decir, no debe malgastar los recursos.

Las tres nociones anteriores conducen a tres tipos de eficiencia: Eficiencia de escala, cuando la organización produce en una escala de tamaño óptimo que es la que permite maximizar el beneficio. Eficiencia asignativa (eficiencia global), cuando la empresa combina los inputs en las proporciones que minimicen el costo de producción y eficiencia técnica, cuando la empresa obtiene el máximo output posible con la combinación de inputs empleada.

Orientación del modelo Orientados a los inputs (recursos, entradas, insumos): Buscan, dado el nivel de outputs (productos, salidas), la máxima reducción en el vector de inputs, es decir, el modelo provee la información, en que tanto están siendo subutiliza-

[1] Aparicio, Baeza Juan. Una Introducción al Análisis Envolvente de Datos. Centro de Investigación Operativa, Universidad Miguel Hernández de Elche.

dos los insumos, mientras permanece en la frontera de posibilidades de producción: Una unidad no es eficiente si es posible disminuir cualquier input sin alterar sus outputs.

Orientados a los outputs: Buscan, dado el nivel de inputs, el máximo incremento de los outputs permaneciendo dentro de la frontera de posibilidades de producción. Este modelo se preocupa por medir que tanto se podría llegar a producir. En este sentido una unidad no puede ser caracterizada como eficiente si es posible incrementar cualquier output sin incrementar ningún input y sin disminuir ningún otro output.

Rendimientos a escala. Rendimientos a escala constante (CCR): Cuando el incremento porcentual del output, es igual al incremento porcentual de los recursos productivos. El modelo CCR proporciona medidas de eficiencia radial, orientados a inputs u outputs. El modelo DEA CCR puede escribirse en términos generales, de tres formas distintas: Fraccional, multiplicativa y envolvente.

Rendimientos a escala variable (BCC): El modelo DEA BCC relaja el supuesto restrictivo del CCR, permitiendo que la tipología de rendimiento a escala que en un momento determinado caracterice la tecnología variable. Este modelo es una extensión del modelo CCR.

Medición de la eficiencia. Para medir la productividad se plantea la siguiente expresión,introducida por Farrel (1957):

$$productividad = \frac{\text{Producción creada}}{\text{Recurso consumido}} = \frac{\text{Salida}}{\text{Entrada}}$$

Por lo general cualquier tipo de unidad u organización que está siendo evaluada (DMU), con el objeto de observar su productividad, tiene más de un input y más de un output, entonces la relación es cambiada por:

$$productividad = \frac{\text{Suma ponderada de salidas}}{\text{Suma ponderada de entradas}}$$

Aquí, se hace necesario el uso de pesos, tanto para cada entrada (v_{rj}) como para cada salida (u_{rj}), con las unidades adecuadas que generan un resultado adimensional, apareciendo el concepto de entrada y salida virtual, como lo menciona (Soto, 2010, pág. 18), así:

$$\text{Entrada Virtual} = \sum_{i=1}^{m} v_{ij} * x_{ij}$$

$$\text{Entrada Virtual} = \sum_{r=1}^{s} u_{ij} * y_{rj}$$

Cuando la productividad de una DMU se compara con la de otras DMU, aparece el concepto de eficiencia relativa, entonces, la eficiencia relativa según Alfredo Roa(2003, pág. 164)la define como:

$$Eficiencia_j = \frac{Eficiencia_j}{Eficiencia_0} = \frac{\text{Salidas Virtuales Entradas Virtuales}_j}{\text{Salidas Virtuales Entradas Virtuales}_0}$$

El subíndice "j"indica la unidad (DMU) que está siendo estudiada (a la que se le va a calcular la eficiencia) y el subíndice "o"la DMU que se toma como referencia. Se pueden distinguir varios tipos de eficiencia relativa en función de la unidad de referencia que se utilice.

Así, la fórmula para calcular la eficiencia relativa es la siguiente:

$$Eficiencia = \frac{\sum_{r=1}^{s} u_{rj} * y_{rj} / \sum_{i=1}^{m} v_{ij} * x_{ij}}{\sum_{r=1}^{s} u_{rj} * y_{ij} / \sum_{i=1}^{m} v_{ij} * x_{ij}}$$

En el anterior cociente, el subíndice "o"al lado derecho de la línea vertical en el denominador nos indica el hecho de que en el denominador se calcula la eficiencia de la DMU que está sirviendo de Referencia. En este cociente se podría encontrar infinitos pesos ponderadores que dan la misma eficiencia. Ver Soto y Arenas (2010) [2].

Con los siguientes pares de pesos v_{ij}, u_{rj} y un múltiplo de ellos, $\alpha * v_{ij}$; $\beta * u_{rj}$, con α, β cualquier número real, se obtiene la misma efciencia.

Para simplificar de alguna forma el número de pesos que dan igual eficiencia relativa se establece de aquí en adelante que la productividad de la unidad de referencia es uno. De esta forma, sea cual fuere la definición de eficiencia relativa utilizada, en el denominador siempre aparecerá la unidad, ya que la unidad de referencia es eficiente, y por lo tanto se puede expresar la eficiencia de DMU "j"como:

$$Eficiencia_j = \frac{\sum_{r=1}^{s} u_{rj} * y_{rj}}{\sum_{i=1}^{m} v_{ij} * x_{ij}}$$

Charnes et al (1978)[3] presentan modelo CCR (Charnes, Cooper y Rhodes) como un salto desde Farrell 1957 a un modelo mejorado y corrigiendo sus fallas, el modelo relacionado a continuación, calcula la eficiencia de una DMU y por lo tanto se necesitan n optimizaciones, una para cada DMU_j : Por ello se asegura en Cooper et al (2007) en la página 23 el sigiente texto traducido al español:

Resolvemos el siguiente problema de programación fraccional para obtener los valores de los "pesos"de las entradas $(v_i)(i = 1, ..., m)$ y los "pesos"de las salidas $(u_r)(r = 1, ..., s)$ como variables.

$$max_{u,v}\theta = \frac{u_1 y_{1o} + u_2 y_{2o} + ... + u_s y_{so}}{v_1 x_{1o} + v_2 x_{2o} + ... + v_m x_{mo}}$$

sujeto a:

$$\frac{u_1 y_{1j} + u_2 y_{2j} + ... + u_s y_{sj}}{v_1 x_{1j} + v_2 x_{2j} + ... + v_m x_{mj}} \leq 1 (j = 1, ..., n)$$

$$v_1, v_2, ..., v_m \geq 0$$

$$u_1, u_2, ..., u_s \geq 0$$

El modelo (1.2) puede también representarse como está en Cook y Seiford (2009)[4] el cual está basado en [3]:

$$max_{u,v}\theta = \frac{\sum_r u_r y_{ro}}{\sum_i v_i x_{io}}$$

Sujeto a:

$$\sum_r u_r y_{rj} - \sum_i v_i x_{ij} \leq 0 \text{ para todo j}$$

$$u_r, v_i \geq \epsilon \text{ para todo r, i.}$$

donde ϵ es un valor no arquimediano designado para forzar la estricta positividad de las variables.

También en Charnes et al (1978) [3] se presenta el modelo de programación no lineal anterior transformando al de programación lineal relacionado a continuación, demostrando mediante 2 teoremas que ambos modelos son equivalentes y que los valores ponderadores para una DMU observada son independientes de las unidades con que las entradas y las salidas son medidas. La eficiencia de una DMU es hallada tras la solución del siguiente modelo de programación lineal:

$$max_{u,v}\theta = u_1 y_{1o} + ... + u_s y_{so}$$

sujeto a:

$$v_i x_{io} + ... + v_m x_{mo} = 1$$

$$u_1 y_{1j} + ... + u_s y_{sj} \leq v_1 x_{ij} + ... + v_m x_{mj}$$

$$(j = 1, ..., n)$$

$$v_1, v_2, ..., v_m \geq 0$$

$$u_1, u_2, ..., u_s \geq 0$$

En Cook y Seiford (2009)[4]: se presenta el modelo no lineal transformado al modelo lineal a continuación.

$$max \sum_r u_r y_{ro}$$

sujeto a:

$$\sum_i v_i x_{io} = 1$$

$$\sum_r u_r y_{rj} - \sum_i v_i x_{ij} \leq 0 \ \forall j$$

$$u_r, v_i \geq \epsilon \text{ para todo r,i.}$$

Continua en Cook y Seiford (2009)[4]

Por dualidad, este problema es equivalente al de programación lineal:

$$min\theta - \epsilon \left(\sum_r s_r^+ \ + \ \sum_i s_i^- \right)$$

sujeto a:

$$\sum_j \lambda_j x_{ij} + s_i^- = \theta_o x_{io}, \ i = 1, ..., m$$

$$\sum_j \lambda_j y_{rj} - s_r^+ = y_{ro}, \ r = 1, ..., s$$

$$\lambda_j, \ s_i^-, \ s_r^+ \geq 0 \ \forall i, j, r$$

θ_o irrestricto

asegurando que el modelo anterior se refiere al "envolvente"o problema primal y que el modelo 1.5 es el problema dual o "multilplier", porque está basado en los pesos ponderadores.

1.3.1. Eficiencia orientada a las entradas o a las salidas. Eficiencia de Pareto. Orientado a las entradas.

Tomado de [2].

Una DMU es Pareto Eficiente si no es posible disminuir ninguno de sus niveles de entrada sin tener que incrementar al menos uno de sus otros niveles de entrada o disminuir al menos uno de sus niveles de salida.
Matemáticamente:

Sean y_{ij} $(r = 1, ..., s)$ los niveles de salida alcanzados por la DMU j y x_{ij} $(i = 1, ..., m)$ los niveles de entradas que ella usa.

Una DMU j_o es pareto eficiente si no existe una DMU $j \neq j_o$ tal que $x_{i'j} < x_{i'j_o}$ para algún $i^{'}$ y $x_{ij \leq x_{ijo}}$ para todo $i \neq i^{'}$ mientras que $y_{rj} \geq y_{rjo}$ para todo r:

Eficiencia de Pareto. Orientado a las salidas.

Tomado de [2]. Una DMU es Pareto Eficiente si no es posible aumentar ninguno de sus niveles de salida sin tener que disminuir al menos uno de sus otros niveles de salida o aumentar al menos uno de sus niveles de entrada.

Matemáticamente:

Sean y_{ij} $(r = 1, ..., s)$ los niveles de salida alcanzados por la DMU j y x_{ij} $(i = 1, ..., m)$ los niveles de entradas que ella usa.

Una DMU j_o es pareto eficiente si no existe una DMU $j \neq j_o$ tal que $y_{r'j} > y_{r'j_o}$ para algún $r^{'}$ y $y_{rj} \geq y_{rjo}$ para todo $r \neq r^{'}$ mientras que $x_{ij} \leq x_{ijo}$ para todo i:

Así de acuerdo con la noción paretiana de eficiencia se considera que una unidad es eficiente si no existe otra en la muestra que produzca más de alguno de los outputs sin producir menos de algún otro y sin utilizar más de alguno de los recursos productivos, o bien, si no existe alguna unidad que produzca los mismos outputs con menos cantidad de algún factor productivo y no más de los restantes.

6.2. Análisis de regresión

El nombre regresión fue introducido por Galton, F. (1800) cuando trató de relacionar las alturas de hijos y padres. Para ampliar mas esta conceptualización se trata el tema de: la estimación de mínimos cuadrados, la cual es una técnica para obtener las estimaciones de los coeficientes β del modelo de regresión lineal. Para el modelo de regresión lineal múltiple, este método consiste en minimizar la suma de cuadrados de los errores $\sum_i \varepsilon_i^2$ con respecto a β, es decir que:

$$\begin{aligned}
\varepsilon^{'}\varepsilon &= (Y - X\beta)^{'}(Y - X\beta) \\
\varepsilon^{'}\varepsilon &= Y^{'}Y - Y^{'}X\beta - YX^{'}\beta^{'} + \beta^{'}X^{'}X\beta \\
\varepsilon^{'}\varepsilon &= Y^{'}Y - 2\beta^{'}X^{'}Y + \beta^{'}X^{'}X\beta
\end{aligned}$$

Usando el hecho de que $\beta^{'}X^{'}Y = (\beta^{'}X^{'}Y)^{'} = Y^{'}X\beta$ y derivando a $\varepsilon^{'}\varepsilon$ con respecto a β e igualando a cero se tiene que:

$$\begin{aligned}
\frac{d\varepsilon^{'}\varepsilon}{d\beta} &= 0 \\
-2X^{'}Y + 2X^{'}X\beta &= 0 \\
-2X^{'}X\beta &= -2X^{'}Y \\
X^{'}X\beta &= X^{'}Y
\end{aligned}$$

La ecuación anterior representa el sistema de ecuaciones normales. Si X es de rango p, $X^{'}X$ es definida positiva y no singular y existe solución única, entonces:

$$\widehat{\beta} = (X^{'}X)^{-1}X^{'}Y$$

Donde $(X^{'}X)^{-1}$ es la inversa generalizada de $X^{'}X$. Entonces,

$$\theta = X\widehat{\beta} = X(X^{'}X)^{-1}X^{'}Y = PY$$

y puesto que P es único, se deduce que P no depende de la inversa generalizada utilizada. Los valores ajustados $X\widehat{\beta}$ se denotan por $\widehat{Y} = (\widehat{Y}_1, \widehat{Y}_2, ..., \widehat{Y}_n)^{'}$ y los elementos del vector $Y - \widehat{Y} = Y - X\widehat{\beta} = (I_n - P)Y$ son llamados residuales y se simbolizan como e. Ahora

el mínimo de $\varepsilon^{'}\varepsilon$ dado por:

$$
\begin{aligned}
e^{'}e &= (Y - X\widehat{\beta})^{'}(Y - X\widehat{\beta}) \\
e^{'}e &= Y^{'}Y - 2\widehat{\beta}^{'}X^{'}Y + \widehat{\beta}^{'}X^{'}X\widehat{\beta} \\
e^{'}e &= Y^{'}Y - \widehat{\beta}^{'}X^{'}Y + \widehat{\beta}^{'}[X^{'}X\widehat{\beta} - X^{'}Y] \\
e^{'}e &= Y^{'}Y - \widehat{\beta}^{'}X^{'}Y \\
e^{'}e &= Y^{'}Y - \widehat{\beta}^{'}X^{'}X\widehat{\beta}
\end{aligned}
$$

es denominado la suma de cuadrados de los residuales (SCR) o suma de cuadrados de los errores (SCE). Como $\widehat{\theta} = X\widehat{\beta}$ es único, observamos que $\widehat{Y}$, e y SCE son únicos independientemente del rango de X.

Las propiedades del estimador de mínimos cuadrados $\widehat{\beta}$, los los valores ajustados $\widehat{Y}$ y los residuales e son todas funciones lineales de las observaciones originales de Y. Si se asume que los ε_i son variables aleatorias independientes con media cero y varianza σ^2, se tiene que $E(\varepsilon) = 0$ y $Var(\varepsilon) = \sigma^2 I_n$. Es de aclarar que:

$$E(Y) = E(X\beta + \varepsilon) = E(X\beta) + E(\varepsilon) = X\beta$$

y

$$Var(Y) = Var(X\beta + \varepsilon) = Var(\varepsilon) = \sigma^2 I_n$$

La $Var(Y)$ es la misma que la $Var(\varepsilon)$, debido a que adicionar una constante como $X\beta$ a una variable aleatoria no cambia la varianza. Por tanto, cuando los errores ε son normalmente distribuidos, Y es también distribuida como una normal multivariada, así que:

$$Y \sim N(X\beta, \sigma^2 I_n)$$

Según lo anterior si el modelo $Y = X\beta + \varepsilon$ es correcto se consideran las siguientes propiedades:

• $E(\widehat{\beta}) = \beta$ Si se asume que los errores son insesgados, es decir $E(\varepsilon) = 0$ y las columnas de X son linealmente independientes, entonces se tiene que:

$$
\begin{aligned}
E(\widehat{\beta}) &= E[(X^{'}X)^{-1}X^{'}Y] \\
E(\widehat{\beta}) &= E[(X^{'}X)^{-1}X^{'}(X\beta + \varepsilon)] \\
E(\widehat{\beta}) &= E[(X^{'}X)^{-1}X^{'}X\beta + (X^{'}X)^{-1}X^{'}\varepsilon] \\
E(\widehat{\beta}) &= E[(X^{'}X)^{-1}X^{'}X\beta] + E[(X^{'}X)^{-1}X^{'}\varepsilon] \\
E(\widehat{\beta}) &= E(\beta) + (X^{'}X)^{-1}X^{'}E(\varepsilon)
\end{aligned}
$$

Como $E(\varepsilon) = 0$, entonces $E(\widehat{\beta}) = \beta$ y por tanto $\widehat{\beta}$ es un estimador insesgado de β.

• $Var(\widehat{\beta}) = \sigma^2(X^{'}X)^{-1}$. Teniendo en cuenta la propiedad de la matriz de varianzas -

covarianzas de Az donde A es una matriz y z es un vector columna, entonces $Var(Az) = AVar(z)A^{'}$. Luego se tiene que:

$$\begin{aligned} Var(\widehat{\beta}) &= Var[(X^{'}X)^{-1}X^{'}Y] \\ Var(\widehat{\beta}) &= (X^{'}X)^{-1}X^{'}Var(Y)((X^{'}X)^{-1}X^{'})^{'} \\ Var(\widehat{\beta}) &= (X^{'}X)^{-1}X^{'}\sigma^2 I_n X(X^{'}X)^{-1} \\ Var(\widehat{\beta}) &= \sigma^2(X^{'}X)^{-1}X^{'}X(X^{'}X)^{-1} \\ Var(\widehat{\beta}) &= \sigma^2(X^{'}X)^{-1} \end{aligned}$$

Por tanto, las varianzas y covarianzas de los coeficientes de regresión estimados están dadas por los elementos de la matriz $(X^{'}X)^{-1}$ multiplicada por σ^2. Los elementos de la diagonal principal son las varianzas en el orden en que los coeficientes de regresión son listados mediante β y los elementos fuera de la diagonal son las covarianzas. Cuando ε es normalmente distribuida, $\widehat{\beta}$ es también distribuido como una normal multivariada, entonces se dice que:

$$\widehat{\beta} \sim N_p(\beta, \sigma^2(X^{'}X)^{-1})$$

De igual manera, al plantear un modelo de estas características con una sola variable independiente y según Walpole, R.(2007) [7] menciona que la forma de relación entre la variable respuesta Y, y las variable independiente es la relación lineal.

$$Y = \alpha + \beta x$$

,

dónde α es la intersección y β la pendiente. Ahora bien, si la relación es exacta, es determinístico entre dos variables y no contiene componentes aleatorios o probabilísticos; por lo tanto, el análisis de regresión identifica la mejor forma de relación entre la variable respuesta y las variables independientes.

Se conoce como regresión a un conjunto de técnicas que se usan para establecer una relación entre una variable dependiente, y una o más variables independientes (predictoras). Dicha relación, se plantea por medio de una ecuación, conocida generalmente como modelo de regresión.

La variable respuesta Y, puede explicarse linealmente con una o mas variables regresoras, lo que indica que puede ser una regresión lineal simple o una regresión lineal múltiple respectivamente.

El modelo de regresión lineal simple de acuerdo a Walpole, R. (2007) [7] "la respuesta Y se relaciona con la variable independiente x a través de la ecuación:

$$Y = \alpha + \beta x + \varepsilon$$

Donde α y β son parámetros desconocidos de la intersección con el eje vertical y la pendiente, respectivamente, y ε es una variable aleatoria que se supone está distribuida normalmente con $E(\varepsilon) = 0$ y $Var(\varepsilon) = \sigma^2$ y donde σ^2 es la varianza del error o varianza residual "(p.391).

Así mismo, el modelo de regresión lineal múltiple según Walpole, R. (2007) [7], está dado por:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$$

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + e_i$$

donde ε_i y e_i corresponden a los errores residuales respectivos, los cuales son asociados con la respuesta y_i, y el valor ajustado estimado.

6.3. Modelo de regresión lineal

6.3.1. Modelo de regresión lineal simple

El modelo de regresión lineal más simple solo involucra una variable independiente y establece que la verdadera media de la variable dependiente cambia a una tasa constante a medida que la variable independiente aumenta o disminuye. De este modo la relación funcional entre la verdadera media de Y_i denotada por $E(Y_i)$, y X_i es la ecuación de una línea recta como es:

$$\mu_i = E(Y_i) = \beta_0 + \beta_1 X_i$$

Donde β_0 es el intercepto, cuando en la anterior ecuación $Y_i = 0$ y μ es el valor esperado de Y_i cuando $X = 0$, y β_1 es la pendiente de la línea, la tasa de cambio en $E(Y_i)$ por unidad de cambio en X. Las observaciones de la variable dependiente Y_i son asumidas a ser observaciones aleatorias de poblaciones de variables aleatorias con la media de cada población dada por $E(Y_i)$. La desviación de una observación de Y_i de su población media $E(Y_i)$ se tiene en cuenta mediante la adición de un error aleatorio ε_i dado al modelo estadístico:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

El subíndice i indica la unidad observacional, es decir (Y_i, X_i) con $i = 1, 2, ..., n$.

Los errores aleatorios ε_i tienen media cero y varianza constante σ^2 y son independientes dos a dos. Los errores aleatorios se suponen normalmente distribuidos, lo cual implica que los Y_i también son distribuidos normalmente. Los errores aleatorios se distribuyen

idéntica e independientemente como una normal con media cero y varianza constante, es decir:

$$\varepsilon_i \sim iidN(0, \sigma^2)$$

6.3.2. Modelo de regresión lineal múltiple

En el análisis de regresión múltiple, el modelo lineal aditivo que relaciona una variable dependiente con k variables independientes es:

$$Y_i = \beta_0 + \beta_{i1}X_{i1} + \beta_{i2}X_{i2} + ... + \beta_p X_{ik} + \varepsilon_i$$

Para el anterior modelo, las observaciones están representadas por (Y_i, X_{i1},Y_{i2},...,Y_{ip}) con i=1,...,n). El subíndice i denota la unidad observacional de los cuales las observaciones de Y y k variables independientes son tomadas y el segundo subíndice designa la variable independiente. El tamaño de muestra es denotado por n con $i = 1, 2, ..., n$, y k denota el número de variables independientes. Hay $(k+1)$ parámetros β_j con $j = 1, ..., k$ a ser estimados cuando el modelo lineal incluye el intercepto β_0. Por notación llamaremos a $k = p + 1$, siendo $n > p$. El modelo lineal en notación matricial:

• Y: Vector columna de observaciones de la variable dependiente Y_i de orden n x 1.

• X: Matriz de orden n X k que contiene una columna de unos, la cual es etiquetada 1 seguida por las k vectores columna de las observaciones de las variables independientes.

• β: Vector de parámetros de orden p X 1 a ser estimados.

• ε: Vector de errores aleatorios de orden n X 1.

Con estas convenciones, el modelo lineal puede ser escrito como:

$$Y = X\beta + \varepsilon$$

o,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{nx1} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}_{nxp} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}_{px1} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{nx1}$$

Cada columna de X contiene los valores particulares de las variables independientes. Los elementos de una fila en particular de X, son coeficientes de los correspondientes parámetros en β.

Los vectores Y y ε son vectores aleatorios. La matriz X es de rango completo. Un modelo para el cual X es de rango columna completo es llamado de modelo de rango completo, β es un vector de constantes desconocidas a ser estimado por medio de los datos.

Al igual que en los modelos de regresión simple los ε_i se asumen que son independientes e idénticamente distribuidos como una normal de variables aleatorias con media cero y varianza σ^2. La función de densidad de probabilidad conjunta de $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ es [8]:

$$\prod_{i=1}^{n} \left[(2\pi)^{-1/2} \sigma^{-1} e^{-\varepsilon_i^2/2\sigma^2} \right] = (2\pi)^{-n/2} \sigma^{-n} e^{-\sum_{i=1}^{n} \varepsilon_i^2/2\sigma^2}$$

Los intervalos de confianza y pruebas de hipótesis estimados de los parámetros están basados en los supuestos de normalidad; aún en ausencia de normalidad, los estimadores de mínimos cuadrados son los mejores estimadores lineales insesgados (MELI), en el sentido de tener mínima varianza entre todos los estimadores lineales insesgados. Si la normalidad se mantiene, los estimadores de máxima verosimilitud se obtienen usando el criterio de búsqueda de aquellos valores de los parámetros que podrían maximizar la probabilidad de obtener una muestra determinada, la cual es llamada función de verosimilitud. Entonces, maximizar la función de verosimilitud en la ecuación (2.7) con respecto a β, es equivalente a minimizar la suma de cuadrados en el exponente y por tanto las estimaciones de mínimos cuadrados coinciden con las estimaciones de máxima verosimilitud [8].

6.4. Regresión Beta

Ésta regresión nos permite modelar una variable respuesta o de interés, con respecto a un conjunto de variables explicativas. De acuerdo a lo expresado por Ferrari, Cribari-Neto (2004) [9], menciona que los modelos de regresión beta permiten precisiones en los coeficientes y estimaciones de parámetros y más exactas, además realiza controles en las asimetrías en la distribución muestral, sin importar su tamaño; adicionalmente, con respecto a los datos pueden ser agrupados y modelados en proporciones, tasas o porcentajes que permite procesos de optimización lineal y no lineal.

Ferrari y Cribari-Neto (2004) [9], propusieron un modelo de regresión, en el que la variable respuesta es continua y está restringida al intervalo (0,1) que se relaciona con otras variables, a través de una estructura de regresión, como se ve a continuación:

$$Y_i|\mu_i, \phi \sim \ Beta(\mu_i, \phi), \ i = 1, 2, 3, ..., n$$

$$g(\mu_i) = \eta_i = \sum_{k-1}^{p} x_{ik}\beta_k, \ p < n$$

donde n es el número de observaciones, p es el número de coeficientes de regresión, g(.)

es una función estrictamente monótona dos veces diferenciables que mapea el intervalo (0,1) en $\mathbb{R}$, $\beta^T = (\beta_1, ..., \beta_p)$ es un vector de coeficientes de regresión $x_{i1}, ..., x_{ip}$ y son observaciones de p covariables, $i = 1, 2, ..., n$.

Una variable aleatoria Y sigue una distribución beta con parámetros $p, q > 0$, denotado por $B(p, q)$, si su función de densidad de probabilidad está dada por

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1},\ 0 < y < 1$$

donde $\Gamma(.)$ es la función gamma. La media y la varianza de Y, son respectivamente,

$$E(Y) = \frac{p}{p+q}\ y\ Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}$$

Si se iguala el

$$E(Y) = \frac{p}{p+q}\ y\ Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}$$

Con el fin de obtener una estructura de regresión para la media, se trabaja con una parametrización diferente de la densidad beta. Siguiendo a Ferrari et al [9], sea $\mu = \frac{p}{p+q}$ y $\delta = p + q$, es decir, $p = \mu\delta$ y $q = (1-\mu)\delta$. Con este cambio de variable, se deduce de la ecuación que

$$E(Y) = \mu\ y\ Var(Y) = \frac{\mu(1-\mu)}{1+\delta}$$

μ es la media de la variable respuesta y δ puede ser interpretado como un párametro de precisión, en el sentido de que, para un μ fijo, un valor grande δ implica un menor valor de la varianza de Y. Para efectos de este trabajo, $\phi = \delta^{-1}$ es un parámetro de dispersión. Bajo esta parametrización, usaremos la notación $Y \sim B(\mu, \delta)$. La densidad de Yse puede reescribir como

$$f(y; \mu, \delta) = \frac{\Gamma(\delta)}{\Gamma(\mu\delta)\Gamma((1-\mu)\delta)} y^{\mu\delta-1}(1-y)^{(1-\mu)\delta-1},\ 0 < y < 1$$

donde $0 < \mu < 1$ y $\delta > 0$.

Por otro parte, si X es una variable aleatoria con distribución binomial de parámetros n y π, entonces la proporción muestral de éxitos es $P = X/n$, y se verifica que n es el número de ensayos y π es la proporción poblacional de éxitos y $X = \sum_{i=1}^{n} X_i$

$$E(P) = \pi\ y\ Var(P) = \frac{\pi(1-\pi)}{n}$$

parámetro μ de la distribución beta con el parámetro π de una variable aleatoria binomial, a la luz de las ecuaciones es claro que

$$Var(Y) = \frac{n}{1+\delta} Var(P)$$

es decir, la varianza de la distribución beta es un múltiplo de la varianza de la proporción muestral. Este hecho permite usar la distribución beta para modelar datos de proporción con sobredispersión, ya que la constante $n/(1+\delta)$ permite modelar la dispersión extra presente en los datos.

6.4.1. Modelo de regresión beta

Sea $Y_1, ..., Y_n$ una muestra aleatoria, tal que $Y_i \sim B(\mu_i, \delta),\ i = 1, ..., n$, es decir, Y_i sigue la densidad dada, con media μ_i y parámetro de precisión δ desconocidos. El modelo de regresión beta supone que la media de Y_i se puede escribir como

$$g(\mu_i) = x_i^t \beta = \eta_i,$$

donde $\beta = (\beta_1, ..., \beta_k)^t$ es un vector de parámetros desconocidos; $x_i = (x_{i1}, ..., x_{ik})^t$ es el vector de k variables regresoras $(k < n)$, las cuales se asumen no aleatorias y conocidas; $\eta_i = \beta_1 x_{i1} + ... + \beta_k x_{ik}$ es el predictor lineal; por lo general $x_{i1} = 1$ para todo i, de manera que el modelo tiene una intersección. Por ultimo, $g(.) : (0, 1) \to \mathbb{R}$ es una función de enlace estrictamente monótona y doblemente diferenciable.

Se supondrá que la respuesta esta limitada al intervalo unitario $(0, 1)$. Sin embargo, el modelo sigue siendo útil para situaciones donde la respuesta es restringida al intervalo (a, b), donde a y b son escalares conocidos, con $a < b$. en ese caso se modela $\frac{(Y_i - a)}{b-a}$ en lugar de Y_i directamente. Además, si Y_i también toma los valores extremos 0 y 1, una transformación útil en la practica es

$$Y^* = \frac{Y(n-1) + 0,5}{n}$$

donde n es el tamaño de la muestra (Smithson & Verkuilen [11]).

Una extensión de la anterior propuesta empleada por Smithson & Verkuilen [11] y formalmente introducida (junto con otras extensiones) por Simas et al [21], es la del modelo de regresión beta con dispersión variable. En este modelo el parámetro de precisión no es constante para todas las observaciones, sino que se modela en forma similar a la media. Concretamente, si $Y_i, ..., Y_n$ son variables aleatorias independientes tales que $Y_i \sim B(\mu_{i,\delta_i}),\ i = 1, ..., n$, los modelos se definen como

$$g_1(\mu_i) = \eta_{1i} = f_1(x_i^t; \beta),$$

$$g_2(\delta_i) = \eta_{2i} = f_2(z_i^t; \Theta)$$

donde, $\beta = (\beta_1, ..., \beta_k)^t$ y $\theta = (\theta_1, ..., \theta_h)^t$ son vectores correspondientes al conjunto de parámetros que se supone funcionalmente independientes; $k + h < n$. η_{1i} y η_{2i} son los predictores, no necesariamente lineales; $x_i^t = (x_{i1}, ..., x_{1_{q1}})$, $z_i^t = (z_{i1}, ..., z_{1_{q2}})$ son, respectivamente, vectores de q_1 y q_2 covariables conocidas.

6.4.2. Inferencia Bayesiana en el modelo de Regresión Beta Univariado

Según Ferreira do Souza, (2011) [12], para estimar los parámetros del modelo propuesto por Ferrari y Cribari-Neto (2004) [9], ellos adoptaron un enfoque frecuentista, para este caso se hace de una manera totalmente Bayesiana en la explicación del modelo y se terminará asignando distribuciones a priori, $p(\beta, \phi)$ para β y ϕ. Para realizar inferencia Bayesiana sobre estos parámetros, se supone que se tiene información sobre las variables aleatorias de $Y_1, ..., Y_n$, cuyos valores pertenecen al intervalo (0,1), el cual es equivalente al intervalo (c,d), transformando las variables en $(Y_1/(d-c), ..., Y_n/(d-c)$. Sea $\beta = (\beta_1, ..., \beta_p)^T$ y $g(\mu_i) = Logit(\mu_i)$, con esta información, se tiene que la densidad de la distribución beta, sobre la reparametrización, del modelo propuesto por Ferrari y Cribari-Neto (2004) [9], para la i-ésima observación queda dado de la siguiente manera:

$$f(y_i|\beta, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_i(\beta)\phi)\Gamma((1-\mu_i(\beta))\phi)} y_i^{\mu_i(\beta)\phi-1}(1-y_i)^{[1-\mu_i(\beta)]_{\phi-1}}$$

$$\alpha \exp\{log\Gamma(\phi) - [log\Gamma(\mu_i(\beta)\phi + log\Gamma((1-\mu_i(\beta))\phi)] + \mu_i(\beta)\phi logit(y_i) + \phi log(1-y_i)\},$$

por consiguiente se tiene que la función de verosimilitud $L(\beta, \phi)$, para n observaciones independientes es:

$$\begin{aligned} L(\beta, \phi) &= \prod_{i=1}^{n} f(y_i|\beta, \phi) \\ &= \Gamma(\phi)^n \prod_{i=1}^{n} \frac{1}{\Gamma(\mu_i(\beta)\phi)\Gamma((1-\mu_i(\beta))\phi)} y_i^{\mu_i(\beta)_{\phi-1}}(1-y_i)^{[1-\mu_i(\beta)]_{\phi-1}} \end{aligned}$$

Entonces, considerando una densidad a priori $p(\beta, \phi)$, se tiene que la densidad a posteriori de estos parámetros es tal que:

$$p(\beta, \phi|y) \alpha\ L(\beta, \phi)p(\beta, \phi)\alpha[\Gamma(\phi)]^n \prod_{i=1}^{n} \frac{1}{\Gamma(\mu_i(\beta)\phi)\Gamma((1-\mu_i(\beta))\phi)} y_i^{\mu_i(\beta)_{\phi-1}}$$

$$(1-y_i)^{[1-\mu_i(\beta)]_{\phi-1}} p(\beta, \phi)$$

Hay diferentes formas para escoger la densidad a priori $p(\beta, \phi)$. Una de esas formas es considerar $p(\beta, \phi) = p(\beta)p(\phi)$, lo que equivale a ϕ y β independientes a priori. Por ejemplo se puede adoptar $\beta_k \sim N(m_k, \Sigma_k^2)$ y $\phi \sim Gamma(a, b)$; pues $\beta_k \epsilon (-\infty, \infty)$, $k = 1, ..., p$ y $\phi > 0$, donde resulta el modelo de regresión beta univariado, dado por:

$$Y_i|\mu_i,\ \phi \sim Beta(\mu_i(\beta), \phi),\ i = 1, 2, 3, ..., n$$

$$g(\mu_i) = \eta_i = \sum_{k=1}^{p} x_{ik}\beta_k,\ p < n$$

$$\beta_k \sim N(m_k, \sigma_k^2)$$

$$\phi \sim Gamma(a, b),$$

con distribuciones a priori para ϕ y $\beta_k,\ k = 1, ..., p$.

Los términos desarrollados $\Gamma(\mu_i(\beta)\phi)$ y $\Gamma(\phi)$ indican que la densidad a posteriori de β y ϕ no poseen forma cerrada. Para generar muestras de esa distribución, se utiliza el método de Monte Carlo vía cadenas de Markov. El uso de las distribuciones a priori mencionadas, llevan a la obtención de distribuciones condicionales completas a posteriori para ϕ y β también de forma desconocida. Por lo tanto un método eficaz para la simulación de generar distribuciones a posteriori es el algoritmo Metrópolis-Hastings. Las distribuciones condicionales completas a posteriori del modelo descrito anteriormente quedan de la siguiente manera:

$$p(\phi|\beta, y) \alpha L(y|\beta, \phi)p(\phi)$$

$$\alpha \exp\left\{ n \log\Gamma(\phi) - \sum_{i=1}^{n} [\log\Gamma(\mu_i(\beta)\phi) + \log((1 - \mu_i(\beta))\phi)] + \right\}$$

$$\phi \sum_{i=1}^{n} [\mu_i(\beta) logit(y_i) + \log(1 - y_i)]\ x\ \phi^{a-1} \exp(-b\phi)$$

y

$$p(\beta_k|\beta_{-k}\phi, y) \alpha\ L(y|\beta, \phi)p(\beta_k)$$

$$\alpha \exp\left\{ n \log\Gamma(\phi) - \sum_{i=1}^{n} [\log\Gamma(\mu_i(\beta)\phi) + \log((1 - \mu_i(\beta))\phi)] \right\}$$

$$+ \left\{ \phi \sum_{i=1}^{n} [\mu_i(\beta) logit(y_i) + \log(1 - y_i)] \right\}\ x$$

$$\exp\left\{ -\frac{1}{2} \sum_{k=1}^{p} \left(\frac{\beta_k - m_k}{\sigma^k} \right)^2 \right\}$$

para $k = 1, ..., p$ donde $\beta_{(-k)}$ es obtenido eliminando el k-ésimo término del vector β.

6.4.3. Algunas Medidas de Diagnóstico o estudio de Influencia local

Según Miyashiro (2008) [13], una etapa importante en el ajuste de un modelo de regresión es el análisis de diagnóstico, que nos permite verificar posibles desviaciones de las hipótesis formuladas para el modelo y nos ayuda a identificar observaciones extremas en algunas diferencias desproporcionales en el resultado del ajuste en un modelo de regresión.

6.4.4. Punto de Apalancamiento

Un posible punto de apalancamiento es aquel que tiene un perfil diferente a los demás en relación a los valores de las variables explicativas. En la práctica para lograr determinar un punto de apalancamiento es construyendo una gráfica de h_{tt} versus índices de observaciones $t, t = 1, 2, 3, ..., n$ un valor grande de h_{tt} comparado con las demás observaciones puede indicar que es un punto de apalancamiento.

6.4.5. Punto Aberrante

Un punto aberrante es aquel que presenta un perfil diferente a las demás observaciones, en relación a los valores de la variable respuesta y también posee un valor bajo en la matriz de proyección H. Para identificar gráficamente un punto aberrante se utiliza el residuo estudentizado versus los índices de las observaciones.

6.4.6. Distancia de Cook

Un punto influyente es aquel que ejerce un peso desproporcional en las estimaciones de los parámetros del modelo, el cual posee un perfil diferente a los demás en relación a los valores de la variable respuesta y presenta un valor alto en la matriz de proyección H. Usualmente se construye un gráfico de la DC_t versus índices t para detectar posibles puntos influyentes.

Según Acuña,2007,

i) la distancia de Cook (Cook, 1977)) [4]: Mide el cambio que ocurriría en el vector $\widehat{\beta}$ de coeficientes estimados de regresión (y por lo tanto en el valor ajustado de la

variable de respuesta) si la i-ésima observación fuera omitida. Se calcula por

$$\begin{aligned} CD_i^2 &= \frac{(\widehat{\beta} - \widehat{\beta}_{(i)})' X' X (\widehat{\beta} - \widehat{\beta}_{(i)})}{ps^2} \\ &= \frac{(\widehat{y} - \widehat{y}_{(i)})' (\widehat{y} - \widehat{y}_{(i)})}{ps^2} \\ &= r_i^{*2} \frac{h_{ii}}{p(1 - h_{ii})} \end{aligned}$$

Notar que si el residual estandarizado es muy grande y si el valor leverage es grande, es decir si la observación está bien alejado en la dirección vertical y horizontal entonces su distancia Cook es bien grande. En general un $CD_i^2 > 1$ indica que la i-ésima observación es potencialmente influencial. Una observación con $CD_i^2 < 0{,}1$ no merece ninguna discusión y si su $CD_i^2 < 0{,}5$ merece un poco de atención. Más específicamente una observación con $CD_i^2 > F(0{,}50, p, n-p)$ es considerado como un valor influencial, la razón es que β cae en un elipsoide de confianza centrado en $\widehat{\beta}$ de radio $F(\alpha, p, n-p)$. Aquí p es el número de coeficientes en el modelo. Sin embargo si todos los CD_i^2 son menores que 1 es mejor plotear los valores CD_i^2 para detectar si hay observaciones con valores grandes comparados con los demás.

ii) DFFITS (Belsley, Kuh, y Welsch, 1980). Es similar a la Distancia Cook, excepto por un factor de escala y el remplazo de la varianza estimada s^2 por $s_{(i)}^2$, la varianza estimada del error excluyendo la i-ésima observación en los cálculos. Más precisamente,

$$DFFITS_i^2 = \frac{(\widehat{y} - \widehat{y}_{(i)})' (\widehat{y} - \widehat{y}_{(i)})}{s_{(i)}^2} = t_i^2 \frac{h_{ii}}{(1 - h_{ii})}$$

Un $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ indica un posible valor influencial. Notar que

$$CD_i^2 = \frac{r_i^2}{p t_i^2} DFFITS_i^2$$

iii) DFBETAS (Belsley, Kuh, y Welsch, 1980). Mide la influencia de la i-ésima observación en cada uno de los coeficientes de regresión. Se calcula por

$$(DFBETAS)_{ji} = \frac{\beta_j - \beta_{j,(i)}}{s_{(i)} \sqrt{c_{jj}}}$$

$(i = 1, ..n, j = 0, ...p)$, donde c_{jj} es el j-ésimo elemento de la diagonal de $(X'X)^{-1}$ Un $|DFBETAS|_{ji} > 2\sqrt{n}$ indica un posible valor influencial.

iv) COVRATIO (Belsley, Kuh, y Welsch, 1980)
Mide el efecto en la variabilidad de los coeficientes de regresión al remover la

i-ésima observación. Se define por

$$COVRATIO_i = \frac{det[s^2_{(i)}(X'_{(i)}X_{(i)})^{-1}]}{det[s^2(X'X)^{-1}]}$$

para $i = 1, ?.n$. Donde $det[A]$ significa el determinante de la matriz A. Usando propiedades de determinantes, se puede obtener la siguiente equivalente fórmula

$$(COVRATIO)_i = \left(\frac{s^2_{(i)}}{s^2}\right)^p \frac{1}{(1-h_{ii})}$$

Si $(COVRATIO)_i > 1+3p/n$ o si $(COVRATIO)_i < 1-3p/n$ entonces la i-ésima observación tiene un valor influencial grande.

6.4.7. Gráfica de Probabilidad Normal

La gráfica de probabilidad normal es una herramienta muy útil para validar el ajuste del modelo. Esta gráfica se basada en los residuos estandarizados. Los puntos que se encuentren fuera de la banda de confianza indican que el ajuste no es apropiado para el modelo. Es válida aún cuando la distribución no sea normal.

6.4.8. Distribuciones G y H de Tukey

Siguiendo a Torres (2001)[14] las familias g y h comprenden una considerable variedad de distribuciones con características especiales en cuanto a la asimetría y elongación, por lo cual resulta de gran utilidad cuando se desea simular datos que provienen de distribuciones con formas diferentes a la distribución normal.

6.4.9. Asimetría

Si Z es una variable normal estándar y g es una constante real, la variable aleatoria $Y_g(Z)$, definida como $Y_g(Z) = (\exp(gz) - 1)g^{-1}$ se dice que tiene la distribución g de Tukey para un valor dado de g. El parámetro g controla la magnitud y la dirección de la asimetría.

6.4.10. Elongación

Si Z es una variable normal estándar y h es una constante real, la variable aleatoria $Y_h(Z)$, definida como $Y_h(Z) = Z\exp(hZ^2/2)$ se dice que tiene la distribución h de Tukey para un valor dado de h. El parámetro h controla la cantidad de la elongación de la distribución. Las distribuciones de las familias h de Tukey son simétricas y su valor esperado es cero.

6.5. Modelos aditivos generalizados (GAM)

Los GLM (Roca, 2003) [15] son una extensión de los modelos lineales que cubren de diversos tipos de datos (familia exponencial). Sin embargo, los datos no siempre se van ajustar a la suposición de linealidad, por lo que en este caso el GLM estará mal especificado, los resultados se desviaran de la superficie respuesta elegida, y las conclusiones obtenidas pueden ser erróneas. En otras palabras existen ocasiones donde la modelización obtenida mediante técnicas clásicas paramétricas resultan demasiado rígidas o incluso inadecuadas para los diferentes problemas de interés lo que hace necesario la aplicación de diferentes modelos mas generales y flexibles que permitan una mejor modelación matemática.

En los últimos años han surgido investigaciones en el campo de la estadística funcional no paramétrica que permiten desarrollos y aplicaciones de modelos generales. Hastie y Tibshirani (1990) [16] proponen evitando la suposición de linealidad, la utilización de modelos lineales generalizados (GAM). Estos modelos extienden a los modelos lineales generalizados (GLM) flexibilidad en regresiones no paramétricas, en este sentido los GAM mantienen interpretabilidad de los GLM al suponer que los efectos son aditivos pero incorporan al mismo tiempo la flexibilidad de los métodos de suavización no paramétricos, ya que el efecto de cada covariable no sigue una forma paramétrica fija, sino que depende de una función totalmente desconocida a la que únicamente se le exige un cierto grado de suavidad para que sea posible su estimación. Explícitamente los GAM vienen dados por

$$\mu_x = H(\beta_0 + f_1(X_1) + ... + f_p(X_p)$$

donde la constante β_0 y las funciones parciales unidimensionales $f_1, ..., f_p$ son desconocidas. Es importante señalar que los GAM evitan el llamado "curse of dimensionality "

7. MÉTODOS DE SOLUCIÓN

7.1. Modelo CCR orientado a las salidas

El Análisis Envolvente de Datos involucra nueve variables de las trece inicialmente definidas (tabla 1), de la cuales se selección como inputs cuatro y outputs cinco en la evaluación de eficiencia, cuyo criterio de selección está orientado a variables estáticas para las entradas y las dinámicas para la salida, lo que lo hace ideal para la medida de eficiencia de las unidades académicas de la UT en su prestación del servicio. Esta metodología del Análisis Envolvente de Datos basada en los métodos de frontera, en los cuales se evalúa la salida respecto a la función de producción establecida en éste tipo de modelos, entendiéndose por tal la relación técnica que transforma los factores en productos; es decir, el máximo nivel de outputs alcanzable con una cierta combinación de inputs, o bien, el mínimo nivel de inputs necesarios para la producción de un cierto nivel de outputs.

Tabla 1: Variables analizadas en el modelo CCR.

No	Variables	Descripción
1	Vivienda propia	Estudiantes que al inicio de su carrera universitaria su núcleo familiar posee vivienda propia
2	Vivienda no propia	Estudiantes que al inicio de su carrera universitaria su núcleo familiar no posee vivienda propia
3	Desertores	Número de desertores por cada una de las Unidades Académicas de la Universidad por periodo académico
4	Docentes de Planta	Docentes de planta que para el periodo de observación 2010 a 2014, poseía cada Unidad Académica
5	Docentes catedráticos	Docentes catedráticos que para el periodo de observación 2010 a 2014, posee cada Unidad Académica
6	Gastos anuales	Ejecución de gastos por cada Unidad Académica en cada vigencia de los periodos de observación 2010 al 2014
7	Grupos de investigación	Grupos clasificados en COLCIENCIAS
8	Semilleros de investigación	Número de semilleros conformados por estudiantes de la Universidad
9	Matriculados	Número total de matriculados en cada periodo académico
10	Graduados	Número total de graduados en cada periodo académico
11	Promedio académico	Promedio obtenido por los estudiantes matriculados en cada periodo académico
12	Promedio de pago matricula	Promedio pagado por los estudiantes en cada periodo académico
13	Monitores	Total de monitores asignados en cada unidad académica

Se tomaron los datos de la variables de entrada y salida de los periodos comprendidos entre 2010 semestre A y 2014 semestre B.

Tabla 2: Variables de entrada y salida para el modelo DEA

VARIABLES DE ENTRADA	VARIABLES DE SALIDA
Docentes de planta	Grupos de investigación
Docentes catedráticos	Semilleros de investigación
Ejecuciones de gastos	Matriculados
Promedio de pago por facultad	Graduados
	Promedio académico por facultad

Figura 1: Resultados obtenidos del modelo CCR orientado a las salida

Periodo	Ciencias	Ciencias de la Educación	Ciencias de la Salud	Ciencias Económicas y Administrativas	Ciencias Humanas y Artes	Agronomía	Forestal	Medicina Veterinaria y Zootecnia	Tecnologías
2010_1	0,999	1	1	1	1	0,992	1	1	1
2010_2	0,999	1	1	1	1	0,999	1	0,999	1
2011_1	0,998	1	1	0,998	1	1	1	1	1
2011_2	0,998	1	1	1	1	0,999	1	1	1
2012_1	0,999	1	1	1	1	1	0,520	1	1
2012_2	0,999	0,999	1	0,999	1	1	1	1	1
2013_1	0,999	1	1	1	1	1,028	1	1	1
2013_2	0,998	1	1	1	1	1	1	1	1
2014_1	0,999	1	1	0,996	1	1	1	1	1
2014_2	0,999	1	1	1	1	1	1	0,999	1

El modelo se formaliza asumiendo que las nueve DMU a ser evaluadas, cada una de las cuales consumen cuatro inputs diferentes para producir cinco outputs también diferentes. La DMUj utiliza un monto de $X_j = x_{ij}$ inputs $(i = 1, ..., 4)$ y produce un monto de $Y_j = Y_{kj}$ productos $(k = 1, ..., 14)$. La matriz $4X90$ de medida del producto es designada por Y, y la $5X90$ de medida de los inputs se designa por X. Se asume además que $x_{ij} \geq 0$ y $y_{kj} \geq 0$, para cada uno de los periodos en las unidades académicas, se obtuvo resultados como se puede evidenciar en la figura 1, muy cercanas a la unidad, lo que lo que refleja eficiencias óptimas para las observaciones significativas.

7.2. Implementación de Variables y Modelos

En el presente trabajo se tuvieron en cuenta cada una de las variables, las cuales son descritas en la siguiente tabla:

Tabla 3: Variables analizadas

Nombre	**Contracción**	**Notación**
Porcentaje de deserción	DES	Y
Gastos unidades académicas	GAS	X_1
Número de graduados	GRA	X_2
Promedio académico	PRO	X_3
Promedio de Pago	PRP	X_4
Número de monitores	MON	X_5

Por otra parte, con la utilización de la herramienta computacional GAM, se obtienen los modelos de regresión beta, los cuales se describen en esta tabla:

Tabla 4: Modelos regresión Beta

Número	**Modelo**	**Enlace media**	**Enlace varianza**
1	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Logit	
2	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Loglog	
3	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Probit	
4	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$		Cauchit$(*)$

* Explicado por variables independientes significativas

Con el fin de determinar el modelo que mejor ajuste los datos, se procedió a utilizar la herramienta GAM. En este tipo de modelos las variables independientes que no se pueden incluir en forma paramétrica, se incluyen de manera no paramétrica, ésta es la razón por la que a este tipo de modelos se les denomina *semiparamétricos*.

Continuando con la secuencia de modelos, la tabla siguiente representa los modelos *semiparamétricos* utilizados. En estos se añade la familia distribucional a la que pertenece la variable respuesta Y.

Tabla 5: Modelos semiparamétricos

Número	**Modelo**	**Familia**	**Enlace media**	**Enlace varianza**
5	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Logit	Logit
6	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Logit	Logit$(*)$
7	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Logit	Logit
8	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Logit	Logit$(*)$
9	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Probit	Logit
10	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Probit	Logit$(*)$
11	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Probit	Logit
12	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Probit	Logit$(*)$
13	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Cauchit	Logit
14	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Cauchit	Logit$(*)$
15	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Cauchit	Logit
16	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Cauchit	Logit$(*)$

En la última tabla el signo $*$ significa que la varianza es explicada por las variables independientes más significativas para la media. Las variables de color rojo ingresan al modelo en forma semiparamétrica.

7.3. Resultados y Discusión

A continuación se presentan las estimaciones para cada uno de los modelos descritos en la sección anterior, en cada casilla el primer número corresponde a la estimación, el valor entre paréntesis corresponde a la desviación estándar de la estimación. Se utiliza la libreria gamlss de R.

Tabla 6: Modelos semiparamétrico ajustado

Parámetro	**Modelo**			
$\boldsymbol{\mu}$	1	2	3	4
β_0	-5.6770 (1.7832)	-1.9860 (0.0642)	-2.9230 (0.0877)	-20.2500 (5.7540)
β_1	-0.0003 (0.0001)	-0.0001 (0.00005)	-0.0001 (0.00007)	-0.0015 (0.0005)
β_2	-0.0046 (0.0014)	-0.0016 (0.0005)	-0.0022 (0.0007)	-0.0190 (0.0039)
β_3	1.1785 (0.4821)	0.0391 (0.0174)	0.0557 (0.0237)	5.5820 (1.5110)
β_4	-0.0018 (0.0008)	-0.0009 (0.0004)	-0.0009 (0.00047)	-0.0057 (0.0027)
β_5	0.0041 (0.0016)	0.0016 (0.0006)	0.0021 (0.0008)	0.0064 (0.0030)
$\boldsymbol{\phi}$	80.49	79.03	79.74	
$\boldsymbol{\gamma_0}$				3.3670 (0.3013)
$\boldsymbol{\gamma_1}$				
$\boldsymbol{\gamma_2}$				0.0124 (0.0045)
$\boldsymbol{\gamma_3}$				
$\boldsymbol{\gamma_4}$				
$\boldsymbol{\gamma_5}$				0.0173 (0.0054)
AIC	-367.3427	-365.8209	-366.5687	-383.6670
BIC	-349.8441	-348.3223	-349.0700	-361.1687

En la tabla anterior, se puede evidenciar que el modelo 4 tiene los criterios de información de *Akaike* (AIC) y de *Bayes* (BIC) más pequeños, por tanto éste podría ser el modelo más apropiado de los estudiados para ajustar los datos. Se puede apreciar también que en el modelo 4, la varianza es explicada por el *Intersecto, número de graduados y número de monitores*.

Con el propósito de verificar y comparar las anteriores estimaciones, se realizó un *Análisis Bayesiano* de la información mediante *WinBUGS*, el cual permite juzgar con un buen grado de aproximación los ajustes obtenidos.

Tabla 7: Estimación de parámetros Bayesianos

Parámetro	**Media**	**Desviación Estándar**	**Mediana**	**Rhat**
β_0	-5.67485	2.25998	-5.40800	1.04313
β_1	-0.00035	0.00018	-0.00035	1.00094
β_2	-0.00485	0.00156	-0.00483	1.00285
β_3	1.18793	0.61621	1.11100	1.05020
β_4	-0.00193	0.00083	-0.00193	1.00700
β_5	0.00419	0.00167	0.00422	1.01185
DESVIO	-363.90000			
DIC	-365.2			

Se puede apreciar en la tabla anterior que las estimación puntual de los parámetros

es algo similar a las del modelo 1, el Rhat entre más cercano sea a 1 significa mejor convergencia, mientras que el *Criterio de Información de Desvío* (DIC) es una estimación del *Error Predictivo Esperado* y cuando éste es menor que el desvío se considera una buena escogencia de los parámetros, lo cual en efecto se cumple.

7.4. Modelos Semiparamétricos

Como se dijo anteriormente, existen variables independientes de tipo contínuo, que mediante procedimientos clásicos resulta casi imposible incluirlas en un modelo paramétrico para que expliquen adecuadamente la variable respuesta. Para lograr esto, dichas variables pueden ser incluídas de manera *no paramétrica*, técnica que tuvo sus orígenes en el cálculo numérico. Para la estimación de los parámetros en este tipo de modelos se utilizó la librería **gamlss** de R. Para detalles teóricos se pueden revisar los textos de *Tibshirani* y *Wood*, así como los artículos de *Stasinopoulus* entre otros. En las siguientes tablas se detallan los resultados del análisis para este tipo de modelos y se evidencian los resultados más importantes del estudio. En estos se tendrá en cuenta sólo la función *Desvío*, el *Criterio de Información de Akaike* (AIC) y el *Criterio de Información de Bayes* (BIC).

Tabla 8: Criterios de información del modelo

Criterio	**Modelo**			
	5	**6**	**7**	**8**
DESVIO	-381.3427	-394.2898	-460.0194	-471.0554
AIC	-367.3427	-376.2898	-404.0194	-411.0554
BIC	-336.0610	-334.0246	-353.7915	-349.8441

Se observa claramente que el modelo 8 posee el menor *Desvío* y el menor AIC, a pesar de no poseer el menor BIC, esto hace que el investigador se incline por el modelo 8.

Tabla 9: Criterios de información del modelo

Criterio	**Modelo**			
	9	**10**	**11**	**12**
DESVIO	-380.5687	-393.1171	-459.8097	-470.2643
AIC	-366.5687	-375.1171	-403.8097	-410.2643
BIC	-335.2700	-333.8151	-352.6189	-349.0700

Se visualiza algo similar a lo de la tabla anterior, el menor *Desvío* y el menor AIC corresponden al modelo 12, el menor BIC lo posee el modelo 11. Esto propone inclinación hacia el modelo 12.

Tabla 10: Criterios de información del modelo

Criterio	**Modelo**			
	13	**14**	**15**	**16**
DESVIO	-386.0918	-401.6619	-460.7914	-480.1864
AIC	-372.0918	-383.6619	-403.8097	-404.7914
BIC	-345.1920	-334.7968	-361.1636	-354.5931

Esta última tabla permite ver algo muy similar a lo de las tablas anteriores, es decir menor *Desvío* y menor AIC corresponden al modelo 16.

Con el propósito de configurar un modelo para los datos, el cual no permita dudas ni conjeturas respecto a los criterios de información y a la vez produzca un reflejo fiel de la realidad, se propusieron cuatro modelos adicionales, en los que se combinan los enlaces *cloglog* y *logit*.

Tabla 11: Modelos con enlace cloglog

Número	**Modelo**	**Familia**	**Enlace media**	**Enlace varianza**
17	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Cloglog	Logit
18	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Cloglog	Logit$(*)$
19	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Cloglog	Logit
20	$Y \sim X_1 + X_2 + X_3 + X_4 + X_5$	Beta	Cloglog	Logit$(*)$

El *Desvío*, AIC, BIC y el criterio de *Hannan Queen* (SBC), para estos últimos modelos se describe a continuación:

Tabla 12: Criterios de información modelo con enlace cloglog

Criterio	**Modelo**			
	17	**18**	**19**	**20**
DESVIO	-381.5553	-394.6182	-460.0582	-471.3355
AIC	-367.5553	-376.6182	-404.0582	-411.3356
BIC	-336.3413	-334.0636	-354.1200	-350.0566
SBC	-350.0566	-354.12	-334.0636	-336.3413

Finalmente y teniendo en cuenta que el menor *Desvío* y el menor AIC corresponden al modelo 20. Es decir el modelo con distribución de variable respuesta Beta, en el que se incluyen de manera semiparamétrica las variables independientes: Gastos (X_1), Promedio Académico (X_3) y Promedio de Pago (X_4), la función de enlace para la media es el *Complemento loglog* y la función de enlace para la varianza es *Logit* y además la varianza se explica linealmente a través de las variables independientes más significativas, es nuestro modelo final.

7.5. Interpretación de Media y Varianza, modelo estimado

La estimación de parámetros para la media en el modelo 20 se resumen a continuación:

Tabla 13: Estimación de parámetros (media) mejor modelo

Parámetro	**Estimación**	**Error Estándar**	T	P	**Significancia**
β_0	-5.6130	1.1030	-5.088	0.0000038	***
β_1	-0.0004	0.00009	-4.2588	0.00007	***
β_2	-0.0017	0.0007	-2.468	0.016440	*
β_3	1.173	0.0275	4.261	0.0007	***
β_4	-0.0022	0.0006	-3.745	0.0004	***
β_5	0.0011	0.0006	2.009	0.049002	*

Se evidencia en la tabla anterior que las tres variables independientes y contínuas: *Gastos*, *Promedio Académico* y *Promedio de Pago* y las cuales fueron incluídas en el modelo de forma *no paramétrica* son altamente significativas. Así la media estimada para la i-ésima unidad se escribe:

$$\hat{\mu_i} = 1 - \exp\{-\exp(\hat{\eta_i})\}$$

En donde la componente sistemática $\hat{\eta_i} = \hat{\beta_0} + \sum_{j=1}^{5} x_{ij}\hat{\beta_j}$ con $i = 1, \ldots, n$.

La misma situación ocurre con la varianza del modelo, la cual es explicada de manera lineal por las variables: *Graduados* (X_2) y *Número de monitores* (X_5), según se presenta en la siguiente tabla:

Tabla 14: Estimación de parámetros (varianza) mejor modelo

Parámetro	**Estimación**	**Error Estándar**	T	P	**Significancia**
γ_0	-1.890333	0.174076	-10.859	0.00000	***
γ_2	-0.004458	0.002132	-2.091	0.0408	*
γ_5	-0.014868	0.003023	-4.918	0.00000	***

Se observa claramente que las dos variables son altamente significativas. La varianza estimada puede ser expresada mediante combinación lineal de las variables independientes que la explican, utilizando la siguiente expresión:

$$\hat{\sigma}^2 = \frac{\exp(\hat{\eta_i})}{1 + \exp(\hat{\eta_i})}$$

en donde $\hat{\eta_i} = \hat{\gamma_0} + (x_{i2}\hat{\gamma_2} + x_{i5}\hat{\gamma_5})$ con $i = 1, \ldots, n$.

7.6. Predicción

Ya establecido el mejor modelo para los datos, se puede proceder a realizar las predicciones mediante el mismo, tanto para las unidades muestrales de los datos como para datos propuesto de manera aleatoria, siempre y cuando se respeten las medidas apropiadas de las variables.

La siguiente tabla permite visualizar de manera numérica la aplicación del modelo final 20 al conjunto de las nueve unidades académicas estudiadas.

Tabla 15: Predicciones mejor modelo

Media		Varianza	
Parámetro	**Valor**	**Parámetro**	**Valor**
β_0	0.003643	γ_0	0.131200
β_1	0.631973	γ_2	0.498880
β_2	0.631495	γ_5	0.496283
β_3	0.960500		
β_4	0.631311		
β_5	0.632525		
Total	0.011689	**Total**	0.129000

Se puede observar que el *Porcentaje medio* estimado de deserción considerando *una unidad* para cada variable independiente es es de 0,011689, mientras que la varianza estimada al considerar *una unidad* para cada variable independiente es de 0,129000, aproximadamente 0,13. Si se ignoraran todas las variables en la media, excepto *Promedio Académico* X_3, se puede observar que el *Porcentaje medio* estimado sería de 0,9605. Aquí se visualiza la alta significancia de las variables independientes contínuas que fueron incluídas de manera *no paramétrica* en el modelo. Se pueden obtener los Porcentajes medios estimados para cada una de las unidades muestreadas.

7.7. Predicción para datos propuestos

Otra manera de utilizar el modelo estimado consiste en proponer datos al azar para obtener el *Porcentaje medio* de deserción pronosticada por el modelo. Supóngase que se desea obtener el *Porcentaje medio esperado* cunado se consideran 50 graduados ($X_2 = 50$), 10 monitores ($X_5 = 10$), 300,000 en gastos ($X_1 = 300{,}000$), que la unidad académica tiene un promedio de 3,6 para sus estudiantes (X_3=3.6) y que el promedio de pago por parte de los estudiantes es de 350,000, ($X_4 = 350{,}000$), mediante el modelo se obtiene un *Porcentaje medio esperado* de 0,1063072.

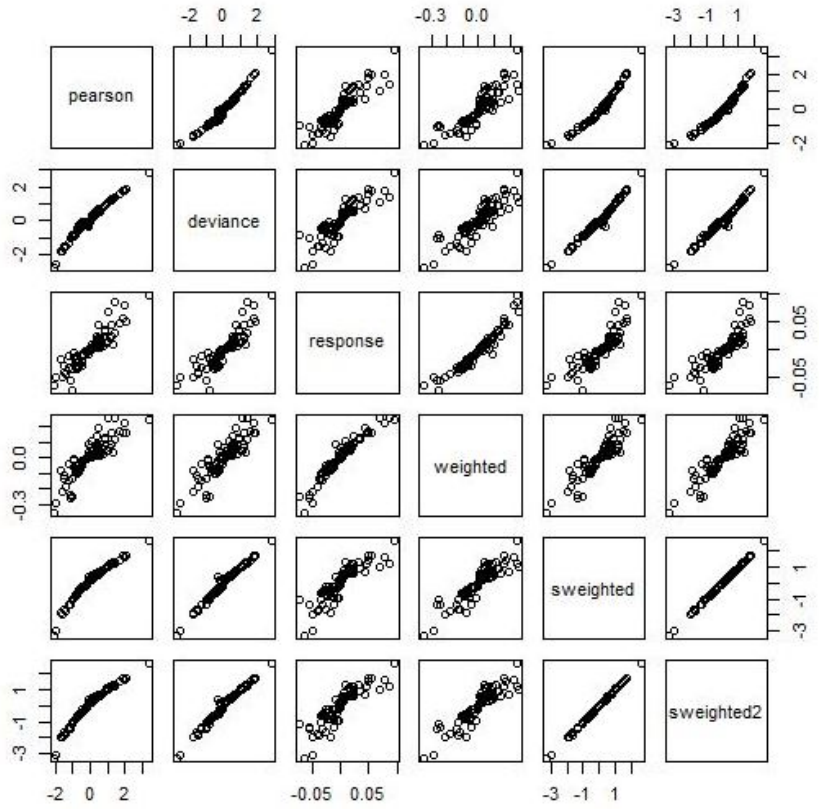


Figura 2: Residuos del modelo estimados

Se puede notar en la figura anterior que hay un comportamiento casi lineal en los residuos lo cual indica que el ajuste del modelo es el adecuado.

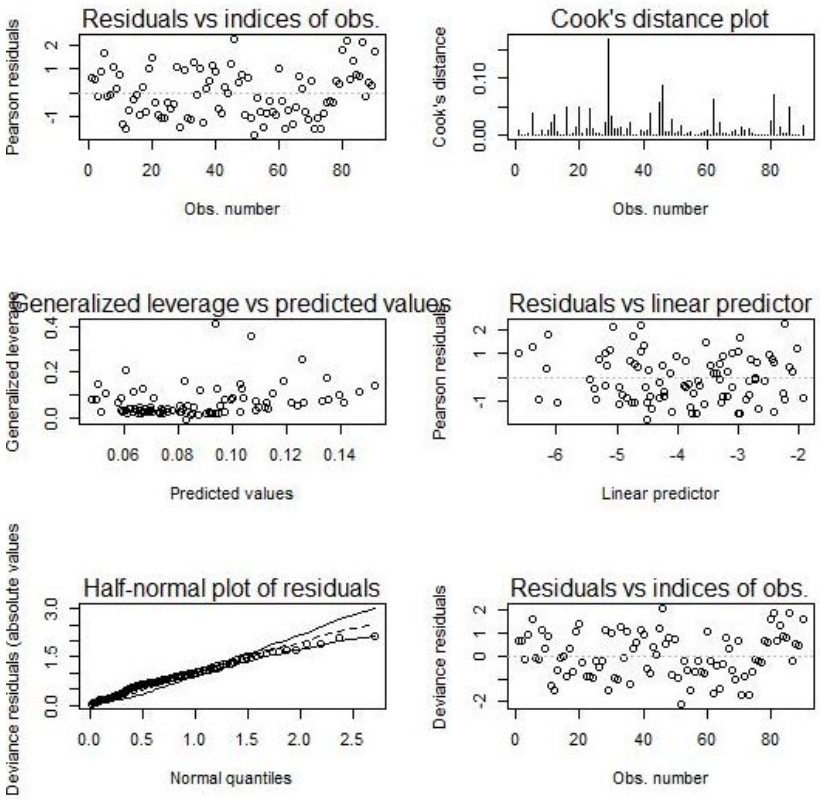


Figura 3: Diagnóstico modelo estimado

Se observa en la gráfica que los residuos están en el intervalo $[-2, 2]$, las distancias de Cook son pequeñas y el ajuste con respecto a los cuantiles de una normal evidencian un buen ajuste.

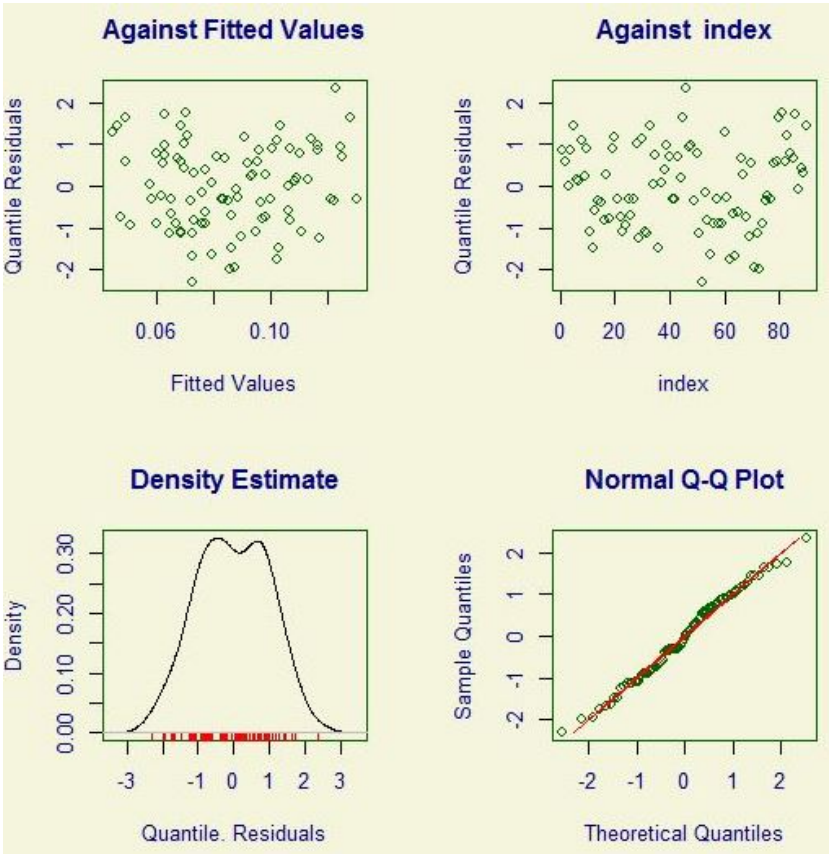
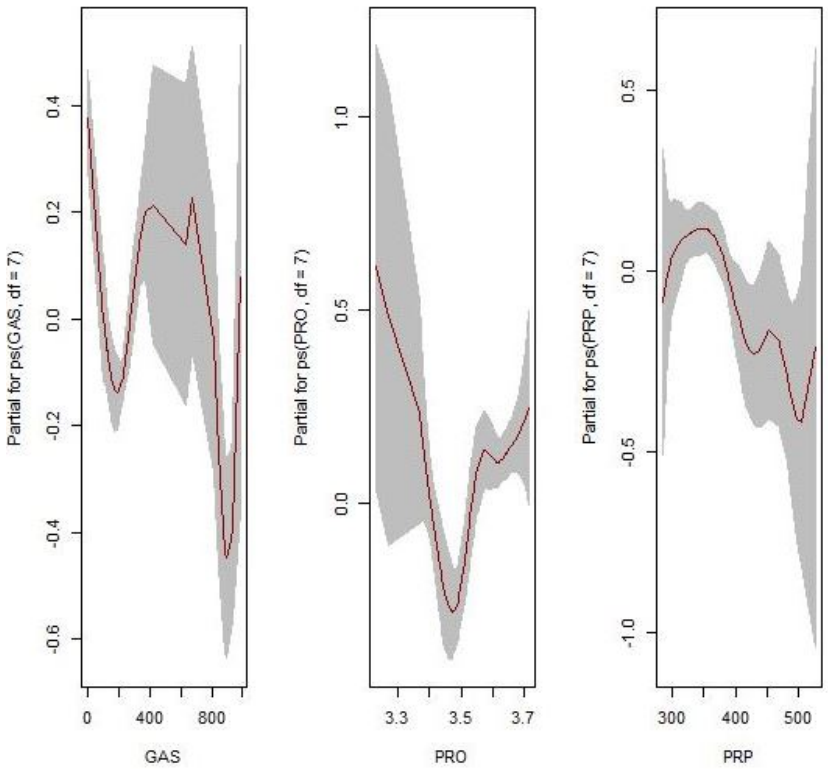


Figura 4: Cuantiles residuales modelo estimado

La gráfica de los cuantiles residuales se encuentran en el intervalo $[-2, 2]$, los cuantiles teóricos versus los cuantiles muestrales tienen comportamiento lineal. Esto permite asegurar un buen ajuste del modelo estimado.



En esta última figura refleja tres momentos en las variables gastos, promedio académico y promedio de pago de matrícula. Los gastos comenzaron una disminución progresiva, se estabilizan en un segundo momento, por último decresen. El promedio disminuye en un comienzo con una alta dispersión, por último aumenta lentamente con poca dispersión. En cuanto a los promedios de pago, se nota una disminución progresiva con una alta dispersión.

7.8. Conclusiones

- Las nueve unidades académicas en su gran mayoría para los 10 periodos evaluados alcanzaron una eficiencia del 100 % en cuanto a los recursos financieros y en general la administración de su infraestructura académico - financiera.

- En el periodo 2012_1, la Facultadad de Ingeniería Forestal presentó una eficiencia del 0.5, lo que implica un seguimiento en la asignación y consecución de recursos acorde a la pontecialidad de esta unidad académica.

- Los Modelos Lineales Generalizados (GLM) son una herramienta poderosa para identificación de patrones estadísticos en datos cuya distribución real pertenece a la familia exponencial. Los modelos clásicos (Regresión y Análisis de Varianza) son sólo casos particulares de GLM, cuando la distribución de los errores es normal.

- Los Modelos Aditivos Generalizados representan una modificación de los GLM, cuando variables independientes contínuas no explican adecuadamente la variable respuesta y deben ingresar al modelo de manera *no paramétrica*, esto es, a manera de funciones o bases que tienen origen en el cálculo numérico.

- Modelos Lineales Generalizados que caracterizan parte *paramétrica* y parte *no paramétrica*, se denominan *Modelos Semiparamétricos*. Estos modelos se identifican con el nombre de *Modelos Aditivos Generalizados de Localización, Escala y Forma*. La herramienta computacional en R, para su análisis se hace a través de la librería: **gamlss**.

- Datos asociados al problema de la deserción en la Universidad del Tolima fueron ajustados mediante **gamlss**, utilizando como densidad de la variable respuesta la *Beta* e incluyendo variables independientes contínuas como predictores.

- Mediante la aplicación de **gamlss**, se puede estudiar simultáneamente la media y la varianza estimadas.

- Se pudo utilizar un análisis *Bayesiano* de los datos como un conductor para la selección del modelo apropiado que ajuste los datos.

- Los modelos semiparamétricos se pueden utilizar como un modelo de fijación de la situación actual de un problema y para realizar pronósticos ó aproximaciones futuras del mismo.

7.9. Recomendaciones

- Se sugiere utilizar modelos DEA categorizados para el tratamiento de cada una de las particularidades de las unidades académicas. **gamlss** en problemas asociados a la deserción, cuando la variable respuesta venga en forma de: proporciones, conteos ó categorías.

- Se propone como alternativa utilizar: *Modelos Aditivos Generalizados Mixtos*, cuando las variables independientes sean difíciles de incluír en el modelo.
- Aunque en el presente trabajo sólo se tuvieron en cuenta 5 variables independientes, se pueden incluír muchas variables de acuerdo a la información provista por las Instituciones.

Apéndice 1.
Código modelo DEA, CCR en Matlab

```
%Entradas
clear all
clc
for j=1:10
%x1 docentes de planta
x1=[1...n];
%x1=log(x1);


%x2 docentes catedráticos
x2=[ 1...n];
%x2=log(x2);

%x3 gastos de cada unidad académica
x3=[ 1...n];
%x4 pagos matricula de los estudiantes, promediado por facultad

x4=[ 1...n];

%x3=log(x3);

%salidas

%y1 grupos de investigación
y1=[ 1...n];

%y2 semilleros de investigación
y2=[ 1...n];

y2=y2/1000000;

%y3 matriculados
y3=[ 1...n];

%y4 graduados de los programas de las nueve unidades académicas
y4=[ 1...n];

%y5 promedios académicos por facultad
y5=[ 1...n];

%ORIENTADO A LAS ENTRADAS
DMUs=9
for k=1:DMUs

%   %ORIENTADO A LAS SALIDAS
f=[0;0;0;0;0;0;0;0;0;-1];
A=[1:j]
```

```
%0 0 0 0 0 0 0 0 0 -1
];
%

b=[x1(k,j);x2(k,j);x3(k,j);x4(k,j);0;0;0;0;0];

%b=[x1(k);x2(k);x3(k);x4(k);0;0;0;0;1];
lb=zeros(10,1);
%[x,fval] = linprog(f,A,b,Aeq,beq,lb)
%[x,fval] = linprog(f,A,b,Aeq,beq,lb)
%x = linprog(f,A,b,Aeq,beq,lb)
x = linprog(f,A,b,[],[],lb)
efi(k,j)=x(10);
end
end
```

Apéndice 2.
Código modelo de regresión Beta en R

```
#########################################################################
### Beta Regression
library(betareg)

data <- read.table("D:/tarea/nubia3.txt",head=T,dec=".",sep="|")
attach(data)
names(data)
#dim(data)
#hist(DES,prob=T)
#lines(density(DES),col=2)

#### Estimation
breg <- betareg(DES ~ DP+SEM+GAS+GRA+PRP | DP+GRI+SEM+MAT+GRA+PRP+MON)
summary(breg)

#### Residuals
des_res <- cbind(
residuals(breg, type = "pearson"),
residuals(breg, type = "deviance"),
residuals(breg, type = "response"),
residuals(breg, type = "weighted"),
residuals(breg, type = "sweighted"),
residuals(breg, type = "sweighted2")
)
colnames(des_res) <- c("pearson", "deviance", "response",
"weighted", "sweighted", "sweighted2")
pairs(des_res)

#### Diagnostics plots
par(mfrow=c(3,2))
plot(breg, which = 1:4, type = "pearson")
plot(breg, which = 5, type = "deviance", sub.caption = "")
plot(breg, which = 1, type = "deviance", sub.caption = "")
which(cooks.distance(breg)>(4/90))

#### Model selection
breg2 <- betareg(DES ~ GAS+GRA+PRO+PRP+MON | GRA+MON, link = "cloglog")
summary(breg2)
breg3 <- betareg(DES ~ GAS+GRA+PRO+PRP+MON | GRA+MON, link = "probit")
summary(breg3)
breg4 <- betareg(DES ~ GAS+GRA+PRO+PRP+MON | GRA+MON, link = "cauchit")
summary(breg4)
par(mfrow=c(3,2))
plot(breg4, which = 1:4, type = "pearson")
plot(breg4, which = 5, type = "deviance", sub.caption = "")
plot(breg4, which = 1, type = "deviance", sub.caption = "")
which(cooks.distance(breg4)>(4/90))
```

```
### Comparing models
AIC(breg,breg2,breg3,breg4)
BIC(breg,breg2,breg3,breg4)

#### Prediction
pred <- predict(breg, type="response")
quant <- predict(breg, type="quantile", at=c(0.25, 0.5, 0.75))
```

Apéndice 3.
Código modelo de regresión Beta Bayesiana en R

```
##############################################################################

#### Bayesian Beta Regression
##############################################################################

library(R2WinBUGS)
library(splines)
library(VGAM)
options("R2WinBUGS.bugs.directory")
folderSNI     <- "D:/tarea/"
setwd(folderSNI)


Y <- DES
n <- length(Y)
X <- cbind(rep(1,n),GAS,GRA,PRO,PRP,MON)
datas1 <- list("Y","X","n")
inits1 <- function(){list(beta=rep(0,6),alpha=rep(0,2))}
parameters1 <- list("beta","alpha")
model1 <- file.path(folderSNI, "modelbeta.txt")
niter <- 100000

resbreg <- bugs(datas1, inits1, n.thin=10, parameters1, model1, n.chains
    =2,
n.iter=niter, n.burnin=20000, bugs.directory="C:/Program Files/
    winbugs14_unrestricted/WinBUGS14")
print(resbreg,5)
names(resbreg)
```

Apéndice 4.
Código modelo regresión Beta Semiparamétrica en R

```
#############################################################################

#### Semiparametric Beta Regression
#############################################################################

library(gamlss)

### logit link function
semibeta11 = gamlss(DES~GRA+MON+GAS+PRO+PRP, family=BE(mu.link="logit",
    sigma.link="logit"))
semibeta22 = gamlss(DES~GRA+MON+GAS+PRO+PRP, sigma.formula=~GRA+MON, family
    =BE(mu.link="logit",sigma.link="logit"))
semibeta33 = gamlss(DES~GRA+MON+ps(GAS, df=7)+ps(PRO, df=7)+ps(PRP, df=7),
    family=BE(mu.link="logit",sigma.link="logit"))
semibeta44 = gamlss(DES~GRA+MON+ps(GAS, df=7)+ps(PRO, df=7)+ps(PRP, df=7),
    sigma.formula=~GRA+MON, family=BE(mu.link="logit",sigma.link="logit"))
cbind(AIC(semibeta11,semibeta22,semibeta33,semibeta44),BIC(semibeta11,
    semibeta22,semibeta33,semibeta44))

### probit link function
semibeta111 = gamlss(DES~GRA+MON+GAS+PRO+PRP, family=BE(mu.link="probit",
    sigma.link="logit"))
semibeta222 = gamlss(DES~GRA+MON+GAS+PRO+PRP, sigma.formula=~GRA+MON,
    family=BE(mu.link="probit",sigma.link="logit"))
semibeta333 = gamlss(DES~GRA+MON+ps(GAS, df=7)+ps(PRO, df=7)+ps(PRP, df=7),
    family=BE(mu.link="probit",sigma.link="logit"))
semibeta444 = gamlss(DES~GRA+MON+ps(GAS, df=7)+ps(PRO, df=7)+ps(PRP, df=7),
    sigma.formula=~GRA+MON, family=BE(mu.link="probit",sigma.link="logit"))
cbind(AIC(semibeta111,semibeta222,semibeta333,semibeta444),BIC(
    semibeta111,semibeta222,semibeta333,semibeta444))

### cauchit link function
semibeta1111 = gamlss(DES~GRA+MON+GAS+PRO+PRP, family=BE(mu.link="cauchit
    ",sigma.link="logit"))
semibeta2222 = gamlss(DES~GRA+MON+GAS+PRO+PRP, sigma.formula=~GRA+MON,
    family=BE(mu.link="cauchit",sigma.link="logit"))
semibeta3333 = gamlss(DES~GRA+MON+ps(GAS, df=7)+ps(PRO, df=7)+ps(PRP, df=7),
    family=BE(mu.link="cauchit",sigma.link="logit"))
semibeta4444 = gamlss(DES~GRA+MON+ps(GAS, df=7)+ps(PRO, df=7)+ps(PRP, df=7),
    sigma.formula=~GRA+MON, family=BE(mu.link="cauchit",sigma.link="logit")
    )
cbind(AIC(semibeta1111,semibeta2222,semibeta3333,semibeta4444),BIC(
    semibeta1111,semibeta2222,semibeta3333,semibeta4444))

### cloglog link function
```

```
semibeta1 = gamlss(DES~GRA+MON+GAS+PRO+PRP, family=BE(mu.link="cloglog",
    sigma.link="logit"))
summary(semibeta1)
plot(semibeta1)

semibeta2 = gamlss(DES~GRA+MON+GAS+PRO+PRP, sigma.formula=~GRA+MON, family=
    BE(mu.link="cloglog",sigma.link="logit"))
summary(semibeta2)
plot(semibeta2)

semibeta3 = gamlss(DES~GRA+MON+ps(GAS, df=7)+ps(PRO, df=7)+ps(PRP, df=7),
    family=BE(mu.link="cloglog",sigma.link="logit"))
summary(semibeta3)
plot(semibeta3)

semibeta4 = gamlss(DES~GRA+MON+ps(GAS, df=7)+ps(PRO, df=7)+ps(PRP, df=7),
    sigma.formula=~GRA+MON, family=BE(mu.link="cloglog",sigma.link="logit")
    )
summary(semibeta4)
plot(semibeta4)

AIC(semibeta1,semibeta2,semibeta3,semibeta4)
BIC(semibeta1,semibeta2,semibeta3,semibeta4)
```

Apéndice 5.
Código predicción y diagnóstico en R

```
##### Predicciones #######
datanew=data.frame(GRA=50,MON=10,GAS=300.000,PRO=3.6,PRP=350.000)
predict(semibeta4,newdata=datanew,type="response",data=data)

#####################################################
### Diagnósticos
#####################################################
plot(semibeta4)
wp(semibeta4)
rqres.plot(semibeta4,howmany=40,plot="average")
dtop(semibeta4)
par(mfrow=c(1,3))
term.plot(semibeta4,terms=3)
term.plot(semibeta4,terms=4)
term.plot(semibeta4,terms=5)
```

Referencias

[1] Universidad del Tolima (2013). *Proyecto Educativo Institucional PEI*; Acuerdo del Consejo Superior N° 022 del 13 de noviembre de 2013. Ibagué.

[2] Soto José A., Arenas V. Wilson. *Análisis Envolvente de Datos de la teoría a la práctica.*Universidad Tecnológica de Pereira, 2010. ISBN: 978-958-44-6403-3, Postergraph S.A, Pereira, Enero 2010.

[3] Charnes A., Cooper W.W, Rhodes E.L. *Measuring the efficiency of decision making units*. European Journal of Operation Research 2, 429-444.

[4] Cook Wade W, Seiford Larry M. (2009) *Data Envelopment Analysis (DEA) - Thirty years on*. European Journal of Operational Research. 192 (2009) 1-17.

[5] Ministerio de Educación Nacional. (2010) *Ingreso, Permanencia y Graduación Boletín informativo (14)*.

[6] Farrell M.J. (1951). *"the Measurement of Productive Eficiency"*. Journal of the Royal Statistical Society Series A, 120 (3), 253-281.

[7] Walpole & Myers (2007) *Probabilidad y estadística para ingeniería y ciencias*. Editorial Pearson.

[8] RAWLINGS, John O., PANTULA, Sastry G., DICKEY, David A. *Applied Regression Analysis*. North Carolina State University. Department of Statistics. Springer texts in statistics. ISBN: 0-387-98454-2, Ed. 2, pp. 75-78, 1998.

[9] FERRARI, S. L. P., CRIBARI-NETO, F. (2004). *Beta regressión for modelling rates and proportions*. Journal of Applied Statistics, 31, 799-815.

[10] Morales, M. (2014). *Prueba de homogeneidad de la dispersión para datos de proporción sobredispersos mediante regresión beta*. Revista Integración. pag:55-70. Recuperado de: http://www.scielo.org.co/pdf/rein/v32n1/v32n1a05.pdf.

[11] Smithson M., Verkuilen J. (2006) *A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables*. Psychological Methods. Vol. 11, No. 1, 54?71. Recuperado de: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.569.7713&rep=rep1 &type=pdf

[12] FERREIRA DO SOUZA, D. (2011). *Regresión Beta Multivariada con Aplicaciones en Pequeñas Areas*. Trabajo de Grado (Instituto de Matematicas). Rio de Janeiro (Brasil) : Universidad Federal de Rio de Janeiro.

[13] MIYASHIRO, E. S. (2008). *Modelos de Regressão Beta e Simplex para Análise de Proporções* . Trabajo de Grado (Instituto de Matemáticas e Estatística). São Paulo: Universidade de São Paulo.

[14] TORRES, J. C. (2001). *Revista Colombiana de Estadística*. Comparación de tres Métodos de Regresión Lineal Usando Procedimientos de Simulación. 24, 33-44. 33.

[15] Roca P. Javier (2003). *Aportaciones a la inferencia no paramétrica en modelos aditivos generalizados y extensiones.*. Universidad de Santiago de Compostela. Addison-Wesley. 1996.

[16] L. Lamport. *falta*. Addison-Wesley. 1996.